International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2390641

Evaluation of the Effectiveness of Feature Selection Methods Combined with Regression Algorithms to Predict Particulate Matter (PM10) in Gandhinagar, Gujarat, India

Zalak L. Thakker[1], Dr. Sanjay H. Buch[2]

[1]Bhagwan Mahavir Centre for Advance Research, Bhagwan Mahavir University, Surat, Gujarat, India

[2]Bhagwan Mahavir College of Computer Application, Bhagwan Mahavir University, Surat, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 01 March 2024

Published: 11 March 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

38-50

ABSTRACT

Feature selection is one of the important data pre-processing techniques that are used to increase the performance of machine learning models, to build faster and more cost-effective algorithms, and to make it easier to interpret the predictions made by the models. The main objective of this research work is to investigate the influence features to predict particulate matter (PM10). This research uses 24-hour average pollutant concentration data of 36 air quality monitoring stations provided by Gandhinagar Smart City Development Limited (GSCDL), Gandhinagar, Gujarat. Important features were identified using five feature selection techniques (correlation, forward selection, backward elimination, Exhaustive Feature Selection (EFS), and feature importance derived using Random Forest Regressor). With selected features six regression algorithms (Multiple Linear Regression, Random Forest, Decision Tree, K-nearest Neighbour, XGBoost, and Support Vector Regressor) were trained to predict PM10. Further, the models were compared based on the Root Mean Square Error (RMSE) and Coefficient of determination (R2) parameters to identify the model with good performance. This proposed model can be utilized as an early warning system, providing air quality information to local authorities to develop air-quality improvement initiatives.

Keywords : Feature Selection, Particulate Matter, Machine learning, Regression algorithm, Gandhinagar, Air Quality, Artificial Neural Network, Decision Tree

I. INTRODUCTION

India is a developing country with rapid economic growth [1]. In line with this progress, there have been numerous significant technological advancements that indirectly contribute to the problem of air pollution. Due to a rise in the population of urban areas and the growth of industries, the levels of air pollution have increased. This has become a major concern for the health of individuals [2]. According to reports released by the United Nations, it is projected that around 60% of the global population will be living in urban regions by the year 2050 [3]. The growth of the population in urban areas will result in an increased need for transportation, energy, and other related essentials. To suffice the need, infrastructure development, transportation services, industrial development as well as residential and

commercial development will be increased in urban areas. This will lead to various pollutions like air, water, noise, etc. Out of these, air pollution is major concern in Indian smart cities. Volatile Organic Compounds (VOCs), Sulfur Dioxide, Carbon Monoxide, Nitrogen Dioxide, Ozone and Particulate Matter are the main air pollutants [4]. In 2019, 1.67 million deaths (17.8% of the total deaths in the India) were due to air pollution. 0.98 million deaths were primarily caused by ambient particulate matter pollution [5]. Particulate Matter (PM) is typically formed in the atmosphere as a result of chemical reactions occurring among various pollutants. The penetration of particles is closely dependent on their size [6]. Particulate Matter (PM) pollution includes PM10 and PM2.5, where PM10 is particulate matter 10 micrometers or less in diameter, PM2.5 is particulate matter 2.5 micrometers or less in diameter. Open burning, power plants, motor vehicle emissions, and industrial process emissions are the major sources of particulate matter pollution. Inhaling tiny liquid or solid droplets, known as particulate matter, can lead to life threatening health problems [7]. The higher concentration of Particulate Matter in environment can cause various serious health issues like cardiovascular diseases and respiratory diseases [8]. Thus, in this hazardous condition, developing improved PM10 prediction models is the greatest answer for regulating particle concentrations, as well as preparing for the worst-case scenarios.

According to [9], features selection techniques can accurately and effectively summarize the original dataset, which then creates new features according to the original dataset. Feature selection process keeps the most important features and their application domain by employing an algorithm or technique. Feature selection improves accuracy and reduces complexity by removing irrelevant features from models. Additionally, it shortens the integration time and creates a simpler model that is considerably simpler to debug [10-12]. The two primary categories of feature-selection methods in machine learning are supervised and unsupervised methods. These two techniques differ in whether its selected features are based on the target variable or not. Supervised method selects features based on the target variable while Unsupervised method selects features without considering the target variable [13]. Feature selection techniques are Wrapper methods, Filter methods and Embedded methods.

Gandhinagar city of Gujarat state is under a smart-city development project. So, due to infrastructure development and industrialization, air pollution across the city have increased. Selection of features varies from dataset to dataset. This study compares the performance of predictive analytics for predicting PM10 based on the optimal number of inputs and influencing factors. In this study we have predicted PM10 exploiting six regression and five feature selection algorithms. Best combination of regression algorithm and feature selection techniques is identified. We have used Root Mean Square Error (RMSE) and Coefficient of Determination (R2) as a performance parameter to compare different models.

II. RELATED WORK

Feature selection methods have been used to predict particulate matter PM10 in several studies. These methods have the objective of choosing the most pertinent characteristics from a dataset to enhance the precision of PM10 Prediction models. Ahmad Zia Ul-Saufie et al. compared different wrapper feature selection methods, including forward selection, backward elimination, stepwise, brute-force, weight-guided, and genetic algorithm evolution, to predict PM10 concentrations in Malaysia. The study found that brute force is the dominant wrapper method in selecting important features for MLR and ANN provides more advantages in terms of model accuracy and feature selection for PM10 prediction [14]. Tina Čok proposed a wrapper-based method called FSBWO, which is based on the Black Widow Optimization technique, and showed improved performance in terms of classification accuracy and selecting optimal subsets of features [15]. Another probabilistic wrapper model was proposed to find relevant features and remove irrelevant ones, improving the predictive accuracy of an induction algorithm [16]. Ani Dijah Rahajoe used a wrapper method called FMF(SES)-GASVM to select features for classification, achieving high accuracy [17]. Bouakline et al.

used the Exhaustive Feature Selection (EFS) method based on statistical scores to select predictors for their deep-learning models [18]. In the paper by Banga et al., five feature selection (Recursive Feature Elimination, Analysis of Variance, Random Forest, Variance Threshold, and Light Gradient Boosting) and six regression algorithms (Extra Tree, Decision Tree, XGBoost, Random Forest, Light GBM, and AdaBoost) were analysed to predict PM2.5 for five cities of China. The AdaBoost algorithm with Light GBM feature selection technique showed the highest performance with the highest performance values (MAE 0.07, RMSE 0.14, and R2 0.94) observed on the Chengdu dataset The computed feature importance revealed that humidity, cbwd, dew point, and pressure play essential roles in air pollution [19]. Sabyasachi in his paper, a new feature selection methodology rooted in particle swarm optimization (PSO) as a wrapper method and an ensemble method was proposed to merge the results of different filter techniques (chi-square, F-regression, and mutual information) to find an optimal feature set that covers most of the key variables of the dataset [20]. Schmainta used the Pearson correlation method to reduce the number of features by observing correlated and anti-correlated attributes to predict PM2.5 and compared the outcomes of the complete set with the reduced set of features [21]. The manuscript introduces a technique for selecting features to forecast air quality. Time series data related to meteorology and air pollution were used for forecasting. The method uses the partial mutual information criterion to select the most informative aggressors for building a prediction model. The selected features are then used to feed an artificial neural network for forecasting PM10 concentration one day in advance [22].

III. METHODS AND MATERIAL

a. Data Collection Process

The five-step data collection process is shown in Figure 1. In the first step, identify the required air quality and meteorological data need to be collected. Locations or places for the data collection are identified in the second stage. In the third stage, an air quality data request was made to Gandhinagar Smart City Development Limited (GSCDL), Gandhinagar, Gujarat, India. The request was processed by GSCDL in the fourth stage and the final fifth-stage data was provided in an Excel file for all requested locations via email.

Figure 1: Steps for data collection process

b. Dataset

The Air Quality dataset of Gandhinagar, Gujarat, India has been considered for the period from 26-01-2022 to 10-07-2023. The dataset is requested from Gandhinagar Smart City Development Limited, Gandhinagar. It consists of meteorological variables and pollutant data from 36 locations in the city. The total number of attributes is 16 (Sr, station, day, month, year, CO, NO2, SO2, Temperature, Humidity, Noise, O3, Uvray, AQI, PM2.5, PM10) and the total records present in each dataset is 13709. The descriptive properties of the dataset are given in Table 1.

Table 1: Dataset Descriptive properties

	Sr.	Station	day	month	year	CO	NO2	SO2	Temperature	Humudity	Noise	O3	Uray	AQI	PM2_5	PM10
count	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00	13709.00
mean	6861.18	18.77	15.61	6.29	2022.48	1.05	43.25	3.78	33.01	58.58	107.80	40.71	5.74	107.25	53.02	80.45
std	3963.66	10.36	8.90	3.44	0.50	2.05	22.30	11.29	4.00	10.26	15.74	21.90	1.68	48.18	33.60	42.22
min	1.00	1.00	1.00	1.00	2022.00	0.00	0.00	0.00	0.13	0.42	15.00	0.00	2.00	2.00	0.00	0.00
0.25	3429.00	10.00	8.00	3.00	2022.00	0.67	30.76	1.43	31.00	52.00	110.00	25.36	4.00	73.00	31.26	50.55
0.50	6859.00	19.00	15.00	6.00	2022.00	0.91	41.23	2.04	33.00	57.00	110.00	35.67	6.00	96.00	48.12	78.50
0.75	10291.00	28.00	23.00	9.00	2023.00	1.20	55.78	2.56	35.00	65.00	116.00	53.22	7.00	122.00	64.02	104.11
max	13736.00	36.00	31.00	12.00	2023.00	135.00	150.00	165.25	46.63	100.00	168.00	150.16	9.00	397.00	594.52	757.08

From Table 1 it was observed that CO, NO2, SO2, O3, PM2.5, and PM10 have minimum values is zero (0) which indicates that missing values are present in those rows. Maximum and average PM10 concentration values are 757.08 and 80.45 respectively. If we compare minimum, 0.25, 0.50, 0.75, and mean values of PM10, PM2.5, O3, AQI, SO2, NO2, and CO, it was observed that all features have outliers present that need to be handled.

c. Data Cleaning

This is a required step while building a regression model which affects the model's performance. The following pre-processing steps have been applied before building the model.

i.Create a single Excel file from multiple worksheets with a station ID field

Air quality data for all 36 locations provided by DSCDL into a separate worksheet. So, it is necessary to combine all location data into a single worksheet to train machine learning models. While combining location data, the station ID field is added to each record to differentiate location data.

ii.Change the datatype of columns from object to float

In the provided dataset Sr. no, station id, and date fields are of numeric type. Other fields such as Co, NO2, SO2, Temperature, Humidity, Noise, O3, Uvray, AQI, PM2.5, and PM10 are of object type so these all fields type need to convert from object type to numeric float type because to train machine learning model all fields need to be of numeric type.

iii.Extract day, month, and year from the date field

From the date field, day, month, and year values were extracted and the date field was removed from the dataset.

iv.Identify and remove '- 'values, null values, duplicate values, and negative values from the dataset.

Gandhinagar air quality dataset contains '-' and null values in some of the rows that were identified and removed from the dataset. The dataset does not contain any negative or duplicate values.

v.Outliers' detection

Outliers present in the dataset affect the machine learning algorithm performance. Boxplot method is used to detect Outliers. Sometimes due to some sudden weather change or due to some activities like construction, heavy traffic, etc. air pollution may increase or decrease. So, if machine learning model will be trained with such type of data outliers it will be able identified uncommon weather condition.

vi.Missing values handling using median imputation

There are missing values present in the provided dataset. Missing values in the dataset are indicated by value zero. Fields CO, NO2, SO2, O3, PM2.5, and PM10 have 93, 7, 599, 24, 76, and 90 missing values respectively. The provided Dataset has outliers present and data are not normally distributed so, the median imputation technique is used to impute missing values.

d. Feature selection techniques

Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in machine learning model building. Techniques for feature selection can reduce overfitting, shorten training times, and enhance model performance. Feature selection can be performed after removing variable redundancy. Three types of variable redundancy are duplicate feature, constant feature, and quasi-constant feature described in Table 2 [23].

Table 2: Types of Variable/Feature Redundancy.

Title	Description
Duplicate feature	Duplicated features are those that in essence are the same. When two features in the dataset show the same value for all the observations, they are in essence the same feature.
Constant feature	Constant features are those that show only one value for all the observations in the dataset (same value for that variable). Variance threshold from sklearn can be used to identify constant features. A simple baseline approach to feature selection where it removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e., features that have the same value in all observations.
Quasi-constant feature	Quasi-constant features are those where a single value is shared by the major observations in the dataset. It's varied but typically, more than 95-99 percent of the observations will present the same value in the dataset. It's up to you to decide the cutoff to call the feature quasi-constant. We are using the variance threshold from sklearn. If the threshold is 0.01, the method will drop the feature if 99% of the observations represent the same value in the dataset.

Feature selection algorithms can be divided into three groups: filter methods, wrapper methods, and embedded methods. Feature selection methods and their subtypes are shown in Table 3.

- **Filter methods** select features based on the intrinsic characteristics of the data, ignoring their interaction with the machine learning model. They are said to be model agnostic.
- **Wrapper methods** are algorithms that wrap the search around a predictive model. They generate multiple feature subsets and then evaluate their performance based on the classification or regression model.
- **Embedded methods** "embed" the selection procedure in the training or induction of the predictive model, provided the optimization process has a way to discriminate among the features.

Table 3: Feature selection methods and its subtypes

Filter Method	Wrapper Method	Embedded Method
Variance Threshold	Forward Feature Selection	Lasso
Correlation	Backward Feature Elimination	Tree-Derived Feature Importance
Chi-Square	Exhaustive Feature Selection	Regression coefficients
ANOVA (Analysis of Variance)		

(1) Filter methods:

Filter methods evaluate the significance of the features by focusing solely on the intrinsic properties of the data. In general, filter methods calculate a feature importance score and then remove features with low scores.

The typical workflow of filter methods involves:

• Ranking features based on some criteria.

• Selecting high-ranking features.

(2) Wrapper method:

Wrapper methods wrap the feature selection procedure around a predictive model. They use the regression or classification model as part of the function that evaluates the subsets' performance.

The idea of wrapper methods is simple: create all possible feature subsets, evaluate those subsets with the machine learning model, and select the best subset. The problem is that for n features, the number of possible subsets is 2^n. Common techniques of the wrapper method are as follows.

A. Forward Feature Selection

In forward selection, variables are progressively incorporated into larger and larger subsets. The algorithms start by training all possible single-variable machine learning models. Then, it selects the feature that returns the best-performing classifier or regression model. In the second step, it creates machine learning models for all combinations of the features from the previous step with all remaining variables in the data. It selects the pair of features that produce the best-performing algorithm. And it continues, adding 1 feature at a time to the feature subset from the previous step until a stopping criterion is met.

B. Backward Feature Elimination

In backward elimination, we start with the set of all variables and progressively eliminate the least promising ones. Backward feature elimination starts by fitting a machine learning model using all the features in the data set and determining its performance. In the next step, it creates all possible feature subsets containing all features except one. It trains a machine learning model for each of the subsets, finds the performance of the model, and then removes the feature that returns the model with the lowest performance. And it continues on and on until a certain stopping criterion is met.

C. **Exhaustive Feature Selection**

The Exhaustive Feature Selection (EFS) algorithm finds the best subset of features out of all possible subsets, according to a performance metric for a certain machine learning algorithm. The procedure will train a machine learning model for each one of the feature subsets with cross-validation and then determine the model's performance.

(3) Embedded methods:

Embedded methods "embed" the selection procedure in the training or induction of the predictive model. In other words, the search for an optimal subset of features is built into the construction of the classifier or the regression algorithm, and as such, embedded approaches train only one machine learning model to select features. A typical embedded feature selection workflow involves the following:

• Training a machine learning model

• Deriving feature importance

• Selecting the top-ranking features

The two main embedded strategies for feature selection are through the Lasso regularization utilized in linear models and through the Feature Importance utilized in decision trees.

e. Performance Metrics

(i) Root mean square error: As stated in Equation 1, this is calculated by taking the square root of the average squared difference between the predicted value and actual value [24].

$$RMSE = \sqrt{\frac{1}{n}(\sum_{k=1}^{n} (yk - ypred)^2)} \quad (1)$$

Where yk is the actual value, ypred is the predicted value, and n is the number of samples.

(ii) R^2: The degree to which the data is statistically near to the regression line is determined in this manner. It has a value between 0 and 1. It describes the variance in the dependent variables that explains the variation in the target variable. R2 is calculated using Equation 2. [24].

$$R^2 = 1 - \frac{SS_{\text{regression}}}{SS_{\text{total}}} \quad (2)$$

Where SSregression is the square sum of regression errors, and SStotal is the squared sum of total error.

IV.METHODOLOGY

The methodology used in our study is shown in Figure 2. In this paper air quality dataset of Gandhinagar City is requested from Gandhinagar Smart City Development Limited, Gandhinagar, Gujarat, India.

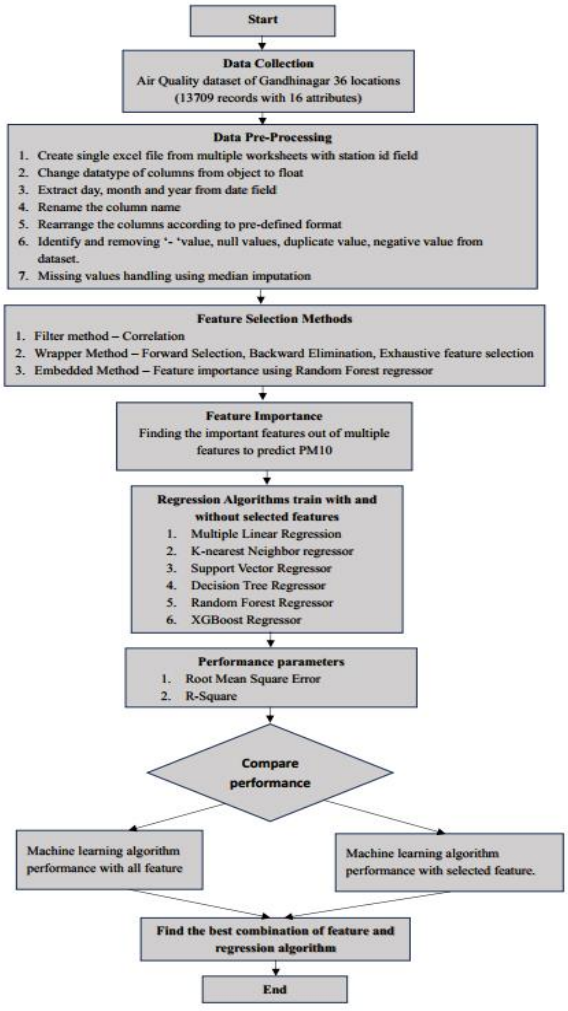

Figure 2. Methodology Flowchart

The data related to air quality was captured over 2 years, from January 2022 to December 2023. The maximum number of attributes in the dataset is sixteen. Firstly, the data was pre-processed to make it suitable for further processing. The dataset contains missing values that we have replaced with median imputation techniques. The outliers in the dataset were identified using a boxplot. From the date column, Day, month, and year features were extracted. Further, we have applied five feature selection techniques (correlation, forward selection, backward elimination, Exhaustive Feature Selection (EFS), and feature importance derived using Random Forest Regressor). The feature selection process may improve performance, reduce training time, and reduce overfitting. There were sixteen attributes in the dataset; Feature selection techniques were applied to find the most influencing features to predict PM10. After that, six regression algorithms such as Multiple Linear Regression, Random Forest, Decision Tree, K-nearest Neighbour, XGBoost, and Support Vector Regressor were applied to predict the PM10 air pollutant. Machine learning regression algorithms were applied first on dataset with all features and second on dataset with selected features using one of the feature selection methods. The results of the both datasets were compared using performance parameters. The performance of all the six regression algorithms was compared based on two performance parameters (RMSE, and R-square) to identify the combination best influencing feature and ML algorithm.

V. RESULTS AND DISCUSSION

Experimental Setup

All the experiments have been performed using the Jupiter version 6.4.8 with python version 3.9.12. on HP Pavilion Gaming laptop 15-ec2xxx with specifications (Windows 10, AMD Ryzen 7 5800H with Radeon Graphics, 3.20 GHz, and 16 GB of RAM). Various Python libraries, such as sci-kit learn, pandas, NumPy, matplotlib, and seaborn, etc., have been used. Data cleaning is done using panda's library. The regression algorithms are applied using the scikit-learn library. All the graphs are plotted using the seaborn/matplotlib library. We conduct an extensive experiment to evaluate the different combinations of five feature selection and six regression algorithms on the dataset of Gandhinagar city of Gujarat. 80% of the total data is considered for training and the remaining 20% for testing.

Variable Redundancy

The Air Quality dataset of Gandhinagar was analysed to identify duplicate features, constant features, and quasi-constant features. Duplicate features/variables are identified using a user-defined function where whereas the Variance threshold from sklearn can be used to identify constant and quasi-constant features. The result shows that the dataset has no duplicate, constant features and quasi-constant features present.

Feature selection

(iii) Filter method

Filter methods do not depend on specific machine learning algorithms. Filter methods are used to remove correlated features from the dataset. The datatype of predictor/input and predicted/output variable is used to identify which filter method will be used for feature selection. Chi-square, ANOVA, and Correlation are three filter methods. Table 4 defines which method will be selected based on the type of input and output feature datatypes.

Table 4: Types of Filter Methods

Chi-Square	ANOVA	Correlation
Categorical Input	Numerical Input	Numerical Input
Categorical Output	Categorical Output	Numerical Output

In air quality dataset of Gandhinagar predictor/ input variable are of Numeric type (Integer and Float) and output/predicted variable are of Numeric type (Integer and Float) so correlation will be used. Figure 3 show the correlation matrix of all features with target variable PM10. Correlation matrix shows the correlation among variables which between the range -1 to 1. Table 5 defines the correlation coefficient value range and its corresponding correlation level.

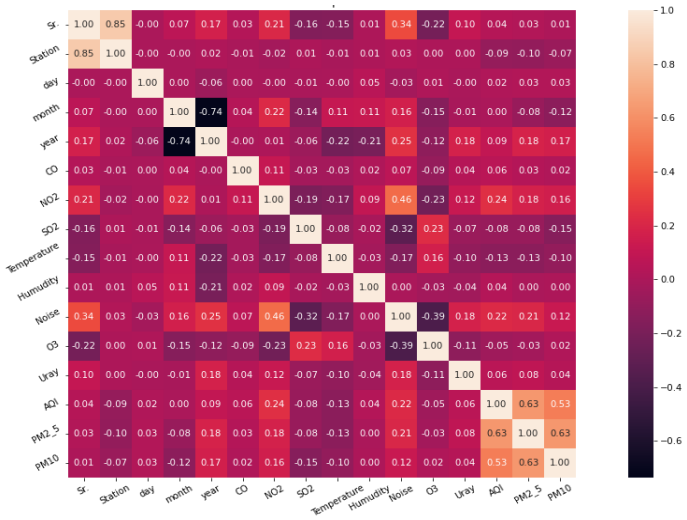

Figure 3: Correlation matrix

Table 5: Range of Correlation Coefficient Values and the Corresponding Correlation Level [25].

Range of Correlation Coefficient Values	Correlation Level	Range of Correlation Coefficient Values	Correlation Level
0.80 to 1.00	Very Strong Positive	-1.00 to -0.80	Very Strong Negative
0.60 to 0.79	Strong Positive	-0.79 to -0.60	Strong Negative
0.40 to 0.59	Moderate Positive	-0.59 to -0.40	Moderate Negative
0.20 to 0.39	Weak Positive	-0.39 to -0.20	Weak Negative
0.00 to 0.19	Very Weak Positive	-0.19 to -0.01	Very Weak Negative

From the above correlation matrix in Figure 3 and the correlation coefficient value range specified in Table 4, it was observed that no features are highly correlated with each other and the target feature also so, all features are used to train machine learning models.

(iv) Wrapper method

Both forward and backward feature selection methods consider adding or removing new features based on their contribution, potentially missing better combinations due to overlooking interactions. So, while using Forward Selection, it May be possible to miss good features that wouldn't be chosen early due to their dependence on other features already included. And for Backward Elimination, it May be possible to remove important features correlated with others, especially if they have weaker individual contributions. Neither method directly addresses collinearity (highly correlated features). Including correlated features can inflate feature importance, leading to misleading

selection. They might miss redundant features that convey the same information as others, reducing model interpretability and efficiency. Exhaustive feature selection has the potential to find the absolute best subset of features for the chosen model and performance metric. It also considers how features influence each other, potentially leading to more accurate models. For the Gandhinagar air quality dataset, the Random Forest regressor gives good accuracy so for the selection of best subset Random Forest regressor was used. Table 6. Shows the selected features using all techniques of wrapper methods.

Table 6: different wrapper feature selection techniques with its selected features.

Wrapper feature selection techniques	Selected features
Forward Feature Selection	'Sr.', 'month', 'year', 'CO', 'NO2', 'SO2', 'Humidity', 'O3', 'AQI', 'PM2.5'
Backward Feature Elimination	'month', 'year', 'CO', 'NO2', 'SO2', 'Humidity', 'O3', 'Uray', 'AQI', 'PM2.5'
Exhaustive Feature Selection	'month', 'year', 'CO', 'NO2', 'SO2', 'O3', 'Uray', 'AQI', 'PM2.5'

(v) Embedded method

Embedded methods search optimal subset of while building machine learning method. Two widely used methods specifically for regression problems are Regularization and feature importance using tree-based algorithm such as decision tree, random forest, etc. Gandhinagar air quality dataset have non-linearity exists and without pre-processing tree-based algorithms gives good accuracy so, to drive important features random forest regressor was used as shown in figure 4. 'Sr.', 'month', 'Noise', 'CO', 'NO2', 'SO2', 'O3', 'Uray', 'Temperature', 'Humidity', 'AQI', 'PM2_5', 'PM10' are important features identified using random forest regressor.

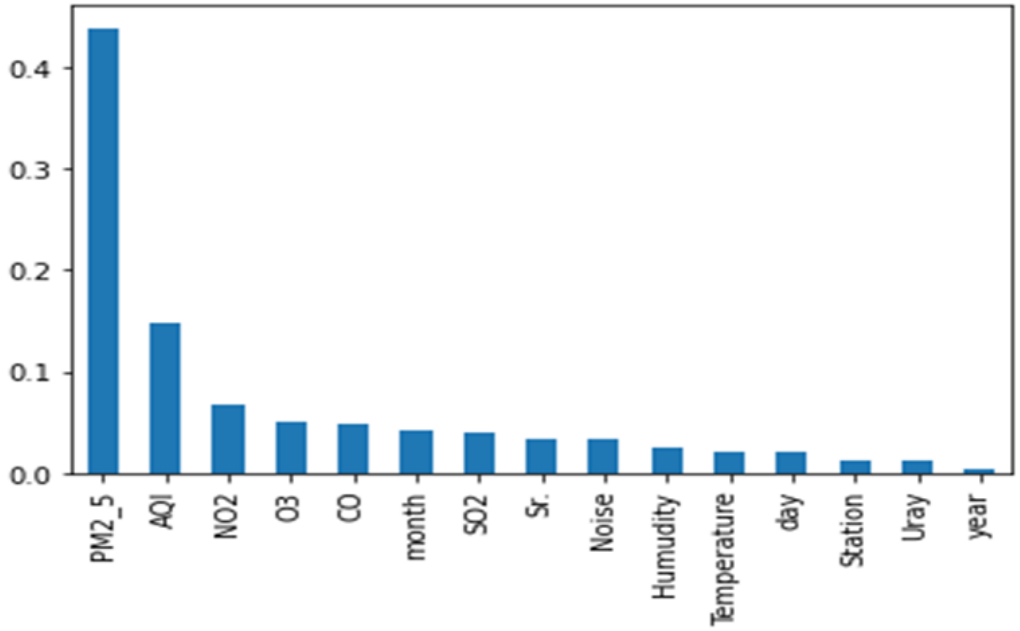

Figure 4 : Feature importance using Random Forest Regressor

f. Machine Learning Algorithm performance before and after Feature Selection

Machine learning algorithms, multiple linear regression, random Forest Regressor, K-nearest Neighbor, Decision tree regressor, Random Forest Regressor, Support vector regressor, and XGBoost Regressor were trained with all the features and with subset of features selected using Wrapper-exhaustive feature selection and Embedded feature selection - Random Forest. The performance of trained model is shown in Table 7.

Table 7: Machine Learning Algorithm performance

PM10 Model	Before Feature Selection		After Feature Selection (Wrapper-exhaustive Feature Selection)		After Feature Selection (Embedded Feature Selection)	
	R2_SCORE	RMSE	R2_SCORE	RMSE	R2_SCORE	RMSE
MLR	48.71	28.99	48.71	28.99	48.71	28.99
RF	77.68	19.12	79.52	18.31	77.60	19.15
KNN	51.85	28.08	48.71	28.99	48.71	28.99
XGB	75.48	20.04	75.39	20.08	77.07	19.38
SVR	1.52	40.16	18.23	36.60	1.57	40.15
DT	51.50	28.19	60.17	25.54	51.38	28.22

g. Discussion

To predict particulate matter (PM10) for Gandhinagar, six machine learning algorithms in combination with five feature selection methods were performed. Using filter-based correlation technique it was observed that features were not highly correlated. 'Sr.', 'month', 'year', 'CO', 'NO2', 'SO2', 'Humidity', 'O3', 'AQI', 'PM2.5' total 10 important features identified using Forward Feature Selection. 'month', 'year', 'CO', 'NO2', 'SO2', 'Humidity', 'O3', 'Uray', 'AQI', 'PM2.5' total 10 important features identified using Backward Feature Elimination. 'month', 'year', 'CO', 'NO2', 'SO2', 'O3', 'Uray', 'AQI', and 'PM2.5' total 9 important features identified using Wrapper - Exhaustive Feature Selection (EFS) technique. 'Sr.', 'month', 'Noise', 'CO', 'NO2', 'SO2', 'O3', 'Uray', 'Temperature', 'Humidity', 'AQI', 'PM2_5', 'PM10' are important features identified using Embedded-random Forest regressor. Forward feature selection and Backward feature elimination ignore collinearity with features other than target variables so sometimes important combinations may be missed. So, important features identified using Exhaustive Feature selection and Embedded Feature selection were used to build six machine learning models. The results indicated that the Random Forest Regressor with features selected by wrapper method – exhaustive feature selection has reported the highest performance in terms of RMSE value 18.31 and R2 value 79.52 on the Gandhinagar air quality dataset. Support Vector Regressor with features selected by wrapper method – exhaustive feature selection has reported poor performance in terms of RMSE value 36.60 and R2 Score 18.23. For using both wrapper and embedded feature selection on the Gandhinagar air quality data set Random Forest Regressor gives good performance and Support Vector Regressor gives poor performance. From the results, it was also identified that tree-based algorithms like Random Forest, Decision Tree, and XGBoost reported good performance for the Gandhinagar air quality dataset.

VI.CONCLUSION

In this paper, Filter, Wrapper, and Embedded feature selection techniques and their subtypes were explored and applied to the Gandhinagar air quality dataset. For the experiment, 36 air quality monitoring stations data were collected from 26-01-2022 to 10-07-2023. Five feature selection techniques (correlation, forward selection, backward

elimination, Exhaustive Feature Selection (EFS), and feature importance derived using Random Forest Regressor) were applied to identify more influencing feature subset out of 15 features. Six machine learning regression algorithms (Multiple Linear Regression, Random Forest, Decision Tree, K-nearest Neighbour, XGBoost, and Support Vector Regressor) were trained with features selected using the Wrapper-Exhaustive Feature Selection (EFS) technique and features selected using Embedded-random Forest regressor to predict the PM10 air pollutant. Random forest regressor with Wrapper - Exhaustive Feature Selection (EFS) technique outperforms compared to other combinations of Machine learning algorithm and feature selection techniques. In the future, various ensemble methods such as stacking, voting, and deep learning algorithms in combination with various Feature selection techniques can be applied to predict air pollutants.

VII. ACKNOWLEDGMENT

The author would like to thank Gandhinagar Smart City Development Limited (GSCDL), Gandhinagar, Gujarat, India for providing Air Quality data and their support.
The authors would like to express their sincere gratitude to Mr. Lovekumar Thakker from Mobile Robotics, Mehsana, Gujarat, India for providing resources, facilities and unwavering support.

VIII. REFERENCES

[1]. Sarabu, Vijay. (2022). INDIA A DEVELOPED COUNTRY?. 10.13140/RG.2.2.15787.72487.
[2]. Zamani Joharestani M, Cao C, Ni X, Bashir B, Talebiesfandarani S
[3]. (2019) PM2.5 Prediction based on random forest, XGBoost, and
[4]. deep learning using multisource remote sensing data. Atmosphere 10(7):373
[5]. Malalgoda C, Amaratunga D, Haigh R (2016) Local governments and
[6]. disaster risk reduction: a conceptual framework. Massey University/The University of Auckland, Auckland
[7]. Manisalidis I, Stavropoulou E, Stavropoulos A and Bezirtzoglou E (2020) Environmental and Health Impacts of Air Pollution: A Review. Front. Public Health 8:14. doi: 10.3389/fpubh.2020.00014
[8]. Pandey, Anamika Brauer, Michael Cropper, et al.," Health and economic impact of air pollution in the states of India: The Global Burden of Disease Study 2019" The Lancet Planetary Health the Lancet Planetary Health volume 5, pages25-38, 2021 DOI: 10.1016/S2542-5196(20)30298-9
[9]. Wilson WE, Suh HH. Fine particles and coarse particles: concentration relationships relevant to epidemiologic studies. J Air Waste Manag Assoc. (1997) 47:1238–49. doi: 10.1080/10473289.1997.10464074
[10]. Cheung K, Daher N, Kam W, Shafer MM, Ning Z, Schauer JJ, et al. Spatial and temporal variation of chemical composition and mass closure of ambient coarse particulate matter (PM10–2.5) in the Los Angeles area. Atmos Environ. (2011) 45:2651–62. doi: 10.1016/j.atmosenv.2011.02.066.
[11]. Hamanaka RB, Mutlu GM. Particulate Matter Air Pollution: Effects on the Cardiovascular System. Front Endocrinol (Lausanne). 2018 Nov 16;9:680. doi: 10.3389/fendo.2018.00680. PMID: 30505291; PMCID: PMC6250783.
[12]. Zhou, H.; Han, S.; Liu, Y. A novel feature selection approach based on document frequency of segmented term frequency. IEEE Access 2018, 6, 53811–53821. [CrossRef]
[13]. Towards Data Science. An Introduction to Feature Selection. 2020. Available online: https://towardsdatascience.com/anintroduction-to-feature-selection-dd72535ecf2b (accessed on 24 February 2024).

[14]. Sukatis, F.F.; Noor, N.M.; Zakaria, N.A.; Ul-Saufie, A.Z.; Suwardi, A. Estimation of missing values In Air Pollution Dataset by Using Various Imputation Methods. Int. J. Conserv. Sci. 2019, 10, 791–804.

[15]. 10. Shaziayani, W.N.; Harun, F.D.; Ul-Saufie, A.Z.; Samsudin, N.; Noor, N.M. Three-Days Ahead Prediction of Daily Maximum Concentrations of PM10 Using Decision Tree Approach. Int. J. Conserv. Sci. 2021, 12, 217–224.

[16]. Libasin, Z.; Suhailah, W.; Fauzi, W.M.; Ul-Saufie, A.Z.; Idris, N.A.; Mazeni, N.A. Evaluation of Single Missing Value Imputation Techniques for Incomplete Air Particulates Matter (PM10) Data in Malaysia. Pertanika J. Sci. Technol. 2021, 29, 3099–3112. [CrossRef]

[17]. Ahmad, Zia, Ul-Saufie., Nurul, Haziqah, Hamzan., Zulaika, Zahari., Wan, Nur, Shaziayani., Norazian, Mohammed, Noor., Mohd, Remy, Rozainy, Mohd, Arif, Zainol., Andrei, Victor, Sandu., Gy., Deák., Petrica, Vizureanu. (2022). Improving Air Pollution Prediction Modelling Using Wrapper Feature Selection. Sustainability, 14(18):11403-11403. doi: 10.3390/su141811403

[18]. Tina, Čok. (2022). Wrapper Based Feature Selection Approach Using Black Widow Optimization Algorithm for Data Classification. doi: 10.1007/978-981-19-3089-8_47

[19]. (2022). Feature Selection and Classification – A Probabilistic Wrapper Approach. doi: 10.1201/9780429332111-72

[20]. Ani, Dijah, Rahajoe. (2019). Forecasting Feature Selection based on Single Exponential Smoothing using Wrapper Method. International Journal of Advanced Computer Science and Applications, doi: 10.14569/IJACSA.2019.0100620

[21]. Oumaima, Bouakline., Y., El, Merabet., Kenza, Khomsi. (2022). Deep-Learning models for daily PM10 forecasts using feature selection and genetic algorithm. doi: 10.1109/ICOA55659.2022.9934503

[22]. Alisha, Banga., Ravinder, Ahuja., Subhash, Chander, Sharma. (2021). Performance analysis of regression algorithms and feature selection techniques to predict PM2.5 in smart cities. International Journal of Systems Assurance Engineering and Management, 1-14. doi: 10.1007/S13198-020-01049-9

[23]. Sabyasachi, Mukherjee. (2022). Ensemble Method of Feature Selection Using Filter and Wrapper Techniques with Evolutionary Learning. 745-755. doi: 10.1007/978-981-19-4052-1_73.

[24]. Stefan, Schmainta. (2023). Correlated Features in Air Pollution Prediction. 527-536. doi: 10.1007/978-981-19-7041-2_44

[25]. Luca, Mesin., Fiammetta, Orione., Riccardo, Taormina., Eros, Pasero. (2010). A feature selection method for air quality forecasting. 6354:489-494. doi: 10.1007/978-3-642-15825-4_66

[26]. Soledad Galli (2022). Feature Selection in Machine Learning with Python, Leanpub

[27]. Banga, Alisha & Ahuja, Ravinder & Sharma, Subhash. (2021). Performance analysis of regression algorithms and feature selection techniques to predict PM2.5 in smart cities. International Journal of System Assurance Engineering and Management. 14. 10.1007/s13198-020-01049-9

[28]. Meghanathan, Natarajan. (2016). Assortativity Analysis of Real-World Network Graphs based on Centrality Metrics. Computer and Information Science. 9. 7. 10.5539/cis.v9n3p7.