International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
OPEN ACCESS
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST52411150

Effect of Feature Scaling Pre-processing Techniques on Machine Learning Algorithms to Predict Particulate Matter Concentration for 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 Feb 2024
Published: 06 Feb 2024

Publication Issue :
Volume 11, Issue 1
January-February-2024
Page Number :
410-419

ABSTRACT

Particulate matter (PM) has widely been recognized as the primary factor responsible for air pollution, posing significant health hazards, particularly cardiovascular and respiratory diseases. Major sources of particulate matter include construction sites, power plants, industries and automobiles, landfills and agriculture, wildfires and brush/waste burning, industrial sources, wind-blown dust from open lands, pollen, and fragments of bacteria. Even though various studies have been carried out to predict particulate matter concentration, there are only a handful of papers that focus on the data scaling pre-processing aspect and how it affects the prediction. For the study, Gandhinagar Smart City Development Limited, Gandhinagar, Gujarat has provided Air Quality data from 26-1-2022 to 16-01-2023. The provided data has several challenges such as missing data, inconsistent data, and mixed data (numerical and categorical). Data pre-processing is an essential step in machine learning regression problems. Data pre-processing techniques include missing value handling, data scaling, outlier detection, feature selection/engineering, and imputation. So, this paper aims to identify the effect of the data scaling pre-processing technique to predict the concentration of Particulate Matter (PM10) for Gandhinagar, Gujarat. Data scaling will be performed based on whether data are normally distributed or not. Four data scaling techniques such as Normalizer, Robust Scaler, Min-Max Scaler, and Standard Scaler in combination with six machine learning algorithms such as Multiple Linear Regressor, Support Vector Regressor, K-Nearest Neighbour regressor, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor were compared to identify best prediction model for Particulate Matter (PM10) concentration.

Keywords : Particulate Matter, Machine learning, Pre-processing Techniques, Data Scaling

I. INTRODUCTION

Air pollution is any substance in the air that harms people, animals, plants or materials. Pollutants come from many sources, both natural and human-made but major contributors are emissions from industry, transport, agriculture and domestic heating [1]. Pollutants include gases, such as sulphur dioxide (SO2), nitrogen oxides (NOx – including nitrogen dioxide [NO2]), volatile organic compounds (VOCs), Ozone (O3) and carbon monoxide (CO), as well as particulate matter (PM) comprised of solid particles and liquid droplets.

Some gases and particles (such as black carbon from partially burned diesel, coal or biomass) are released directly into the air and are therefore called primary pollutants. Secondary pollutants, both gases and particles, also arise from reactions between chemicals in the air [2]. For example, Ozone, sulphate, nitrate and ammonium nitrate can form using chemical reaction process. For air-quality purposes, PM is defined by the size of its constituent parts rather than its chemical composition. PM10 refers to particles up to ten micrometres (μm) across, while PM2.5, describes particles no bigger than 2.5 μm (about 30 times smaller than the width of a human hair) [3]. Particulate matter" (PM) refers to solid particles and liquid droplets found in the air. These airborne particles and droplets vary in composition and size [4]. The prediction of Particulate Matter from available air quality data helps to take required action for controlling PM concentration in the environment. Air quality data including meteorological and pollutants, are recorded at monitoring station locations. Data collected from monitoring stations are not always in the form that directly available for prediction algorithm. Data preparation is an integral part of the data mining process. It refers to the process of cleaning, converting, and combining data prior to feeding data for prediction. The purpose of data preparation is to increase the data's quality and suitability for the machine learning algorithm. Data preparation steps include data cleaning, data integration, data transformation, data reduction, data discretization, data normalization or data scaling. Data cleaning involves missing value handling, outliers detection and removal, duplicate data identification for identifying and correcting errors in data. Data integration step combine data from various sources to produce a single dataset. Data integration can be tricky as it involves handling data in many forms, structures, and semantics. Normalization, standardization, and discretization are three common data transformation methods. Normalization scales down data within the range (for example 0 to 1), whereas standardization transforms data to have mean value of zero with a standard deviation of one. Continuous data will be converted into discrete categories using discretization process. feature selection and feature extraction techniques are used to reduce the size of the dataset while preserving the important information.

In this paper, different feature scaling pre-processing techniques identified from the previous study and applied to predict particulate matter. Feature scaling was applied on numeric features in the dataset. Feature scaling methods scales down numeric feature values. For the study data were collected from Gandhinagar Smart City Development Limited, Gandhinagar, Gujarat, India. Six machine learning algorithms were trained with and without applying feature scaling on dataset and performance was compared.

II. METHODOLOGY

1.1 Study Area and data collection

There are in total 36 air quality monitoring stations within Gandhinagar city, Gujarat, India are being covered in this study. Data are collected from Gandhinagar Smart City Development Limited (GSCDL). GSCDL provides 24 Hour average data from 26-1-2022 to 16-01-2023 for all 36 locations. Table 1 shows the selected location names.

Sr. No.	Monitoring Station	Sr. No.	Monitoring Station
1	Akshardham	**19**	Indroda
2	CH-0 Circle	**20**	Infocity
3	CH-3	**21**	K-5
4	CH-5	**22**	K-6
5	CH-6	**23**	K-7
6	CHH4A	**24**	KH-0(Sargasan)
7	CHH-6	**25**	KH-1
8	Civil Hospital	**26**	KH-2
9	G-1	**27**	KH-5
10	G-3	**28**	KH-6
11	G-5	**29**	Old Sachivalay Main Gate
12	G-6	**30**	Reliance Chokadi
13	G-7	**31**	Sarita Udhyan
14	GH-1	**32**	Sector 22 shopping mall
15	GH-2	**33**	Sector 6 Shopping Centre
16	GH-3	**34**	Sector-1 Garden
17	GH-5	**35**	Sector-28 Garden
18	GH-6	**36**	Sector-7 shopping mall

Table 1 : Air Quality Monitoring Station Locations

Total 7464 records were received across all 36 monitoring stations. Figure 2 shows how the dataset was randomly divided into 80% for training and 20% for test sets to predict PM10.

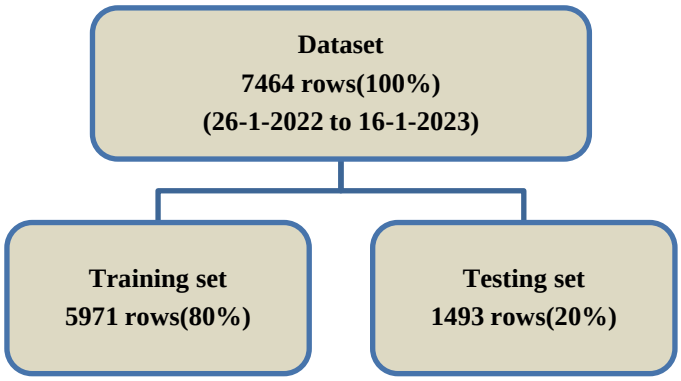

Figure 2 : Train test Split of dataset

The dataset comprises meteorological fields such as temperature, humidity, noise, UV Ray, and pollutants including CO, NO2, SO2, O3, PM2.5, PM10, and AQI. To predict particulate matter list of predictors and predicted features are described in Table 2.

Features	PM10 Prediction	PM2.5 Prediction
predictor Features	Sr Station day month year CO NO2 SO2 Temperature Humidity Noise O3 Uray AQI PM2_5	Sr Station day month year CO NO2 SO2 Temperature Humidity Noise O3 Uray AQI PM10
predicted Feature	PM10	PM2_5

Table 2 : Predictor and Predicted Features for Prediction Models

1.2Feature Scaling

Many machine learning algorithms perform better when numerical input variables are scaled to a standard range. This includes algorithms that use a weighted sum of the input, like linear regression, and algorithms that use distance measures, like k-nearest neighbours. The two most popular techniques for scaling numerical data before modelling are normalization and standardization. Normalization scales each input variable separately to the range 0-1, which is the range for floating-point values where we have the most precision. Standardization scales each input variable separately by subtracting the mean (called centering) and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one [5].

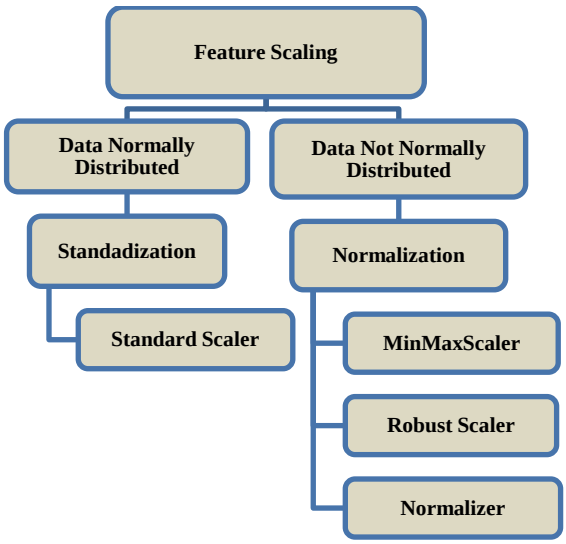

Figure 7 : Feature Scaling Methods

(1) **Standard scaler** or more commonly known as standardization techniques transform the dataset to have a mean value of zero with a standard deviation of one. The transformed value or often called z-score is calculated by [6]:

$$z = \frac{x - \mu}{\sigma}$$

where, x is the original value, μ is the mean value, and σ is the standard deviation. Standardization could be useful as most machine learning estimator prefer to work with features that looks like a normally distributed data.

(2) The **min–max scaler** converts the numerical values v in the dataset of a numerical attributes A to a given range R of [new–min$_A$, new–max$_A$]. The transformed values for this feature scaling technique will falls between the interval of 0 and 1 [6]. The formula to calculate the transformed value v is depicted as follow:

$$v' = \text{new} - \min_A + R \cdot \frac{v - \min_A}{\max_A - \min_A}$$

where max$_A$ and min$_A$ are the original maximum and minimum range of the numerical attributes, respectively.

(3) **Robust scaler** eliminates the median value and scale the data based on the interquartile range that has a span between 1st quartile (25th quantile) and the 3rd quartile (75th quantile). The transformed value for robust scaler falls between –2 and 3, which are relatively higher than the other scaling methods. The formula to calculate the transformed values for robust scaler is [6]:

$$v' = \frac{v - Q1_A}{Q3_A - Q1_A}$$

where v is the original value, A is the numerical attribute, and Q3A – Q1A is the interquartile range.

(4) **Normalizer** scales each sample in the dataset instead of the whole parameter into a unit of normalized value. For instance, if the feature has a value of x, y, and z then the transformed values v can be calculated by this formula [6]:

$$v' = \frac{x_i}{\sqrt{x_i^2 + y_i^2 + z_i^2}}$$

1.3 Accuracy Measure

To compare machine learning algorithm

performance two accuracy measures were used namely Root Mean Squared Error (RMSE) and Coefficient of determination (R^2).

1.3.1 RMSE

Root mean square error measures the square root of the mean of the squared differences between predicted and actual values. It is calculated as [7]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}}$$

where n is the number of records, P_i and A_i are the predicted and actual values, respectively. RMSE has the same unit of measurement of the actual or predicted values which is $\mu g/m^3$. When RMSE has less value, it means the model predicts better.

1.3.2 R^2 Score

R^2(Coefficient of determination) evaluates the correlation between actual and predicted values. It is calculated as [7]:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(A_i - \bar{A})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2 \sum_{i=1}^{n}(P_i - \bar{P})^2}} \right)^2$$

where n is the number of records, A& and P& are the actual and predicted values, respectively. $\bar{A}$ and Pc represent the mean measured and mean predicted value of the pollutant, respectively. RY is a descriptive statistical index which has no dimensions and ranges from 0 to 1 corresponding to no correlation and complete correlation, respectively.

1.4 Experimental Setup

Six machine learning algorithms, Multiple Linear Regression, Random Forest, Decision Tree, K-nearest Neighbour, XGBoost, Support Vector Regressor were chosen to compare the overall performance of different feature scaling pre-processing techniques. The implementation of all six ML algorithms and study results took place 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). This study uses built-in preprocessing libraries provided by Scikit-learn tools: Normalization, Standardization, Min-Max Scale, Standard scale, Robust Scaler. Figure 4 illustrates the overall experimental approach using a flowchart.

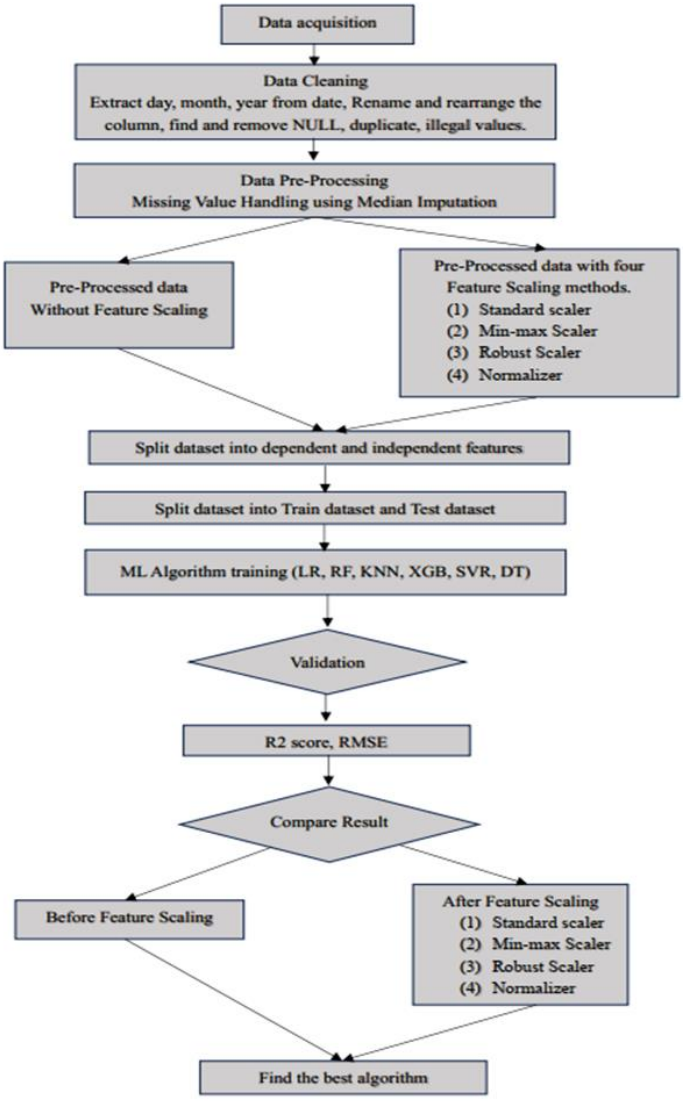

Figure 4 Flowchart

III. RESULTS AND DISCUSSION

3.1 Data Distribution

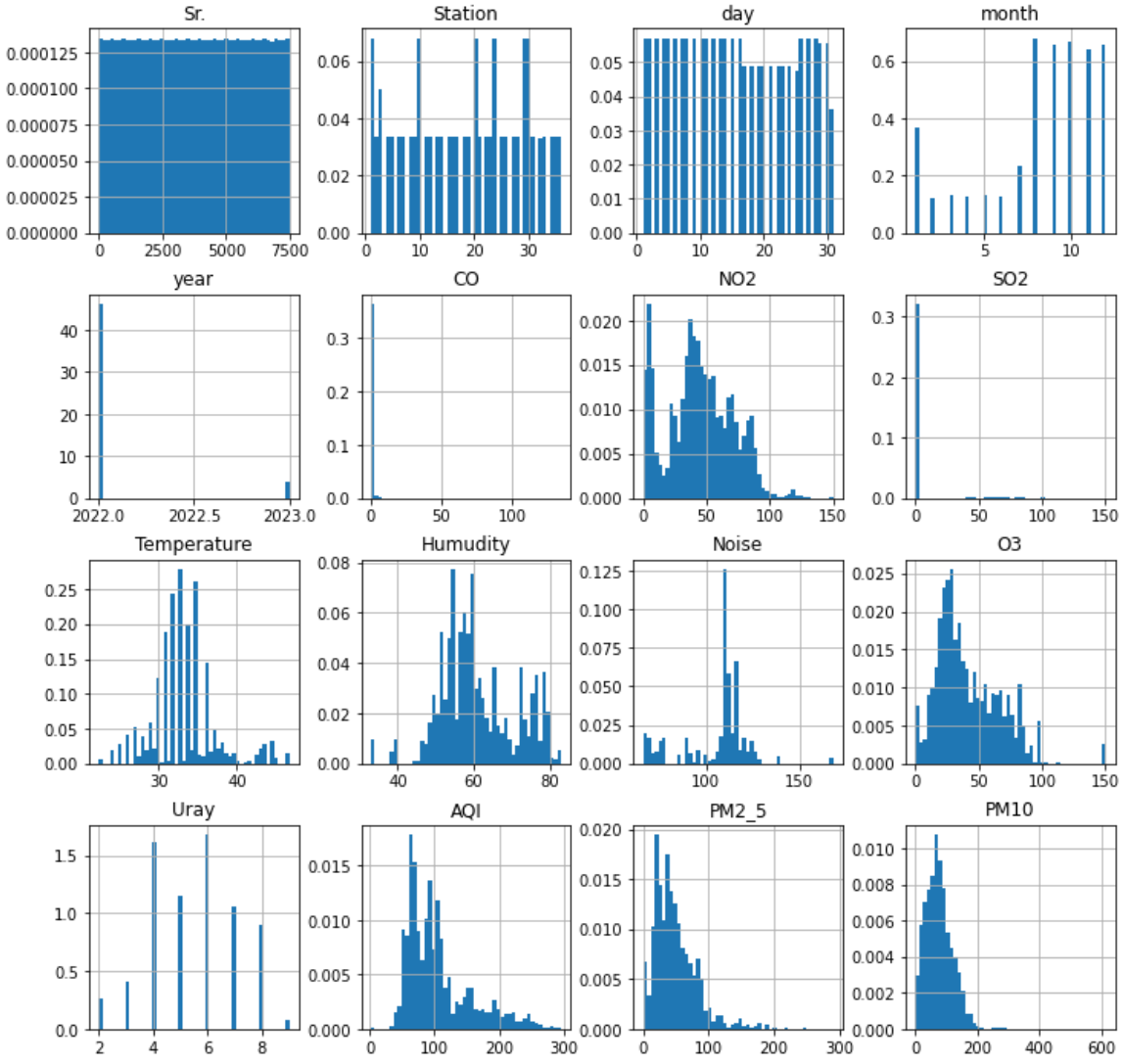

Figure 6 : Dataset Feature Distribution

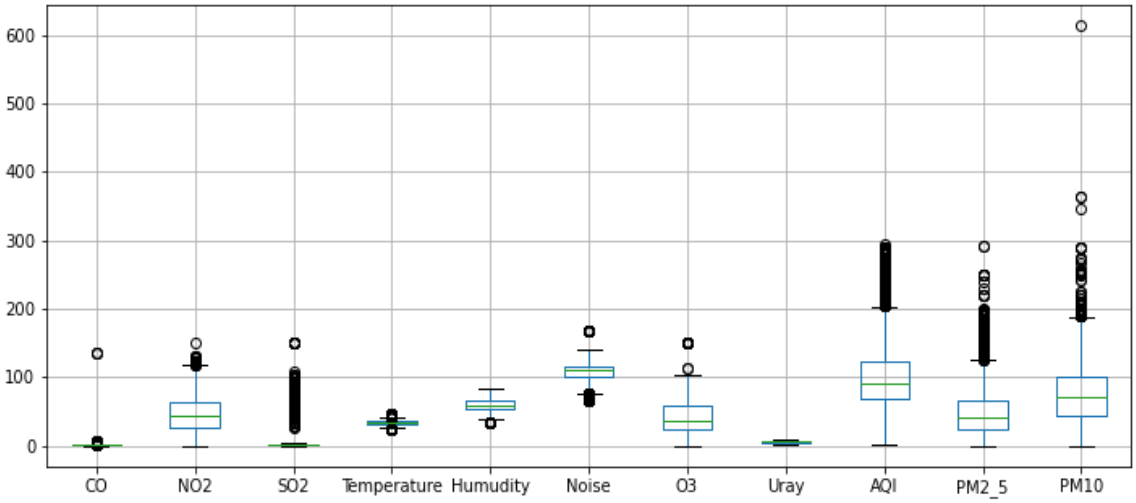

Figure 7 : Boxplot diagram for Outliers in the dataset.

By analysis of the considered dataset, it is observed that the dataset has outliers (as shown in Figure: 7) and the distribution of data (as shown in Figure: 6) is not normal.

1.5 Prediction model without pre-processing.

The Machine Learning Algorithms are Multiple Linear Regression (LR), Random Forest (RF), K nearest Neighbour(KNN), XGBoost (XGB), Support Vector Regressor(SVR), Decision Tree Regressor(DT) were initially used to train models with the original air quality data of Gandhinagar city. The models are trained without applying any pre-processing steps. The accuracy of the resulting PM10 prediction model can be observed from table 3, where the results are representative of the test accuracy. Note that the default value of the missing data is set to zero.

PM10	Random_state=None		Random_state=None		Random_state=None		random_state=0		random_state=20		Random_state=42		random_state=125	
Model	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
LR	37.76	33.20	35.63	33.53	40.36	32.48	41.00	33.40	37.90	36.87	35.58	35.45	35.19	33.56
RF	69.47	23.25	73.41	21.55	72.15	22.19	71.97	23.00	70.46	25.39	66.51	25.58	69.83	23.06
KNN	41.31	32.24	37.21	33.11	47.73	30.40	50.69	30.54	50.99	32.76	44.49	32.90	40.52	32.15
XGB	73.34	21.73	75.11	20.85	75.59	20.78	74.58	21.92	72.35	24.60	69.52	24.38	72.27	21.95
SVR	1.40	41.79	1.33	41.51	1.51	41.74	-0.29	43.55	1.87	46.35	1.94	43.73	1.33	41.41
DT	40.50	32.46	33.29	34.13	49.33	29.94	43.48	32.70	44.07	34.99	42.24	33.56	31.62	34.47

Table 3 : R2 and RMSE values for different PM10 regression model with different Random State value in train test split

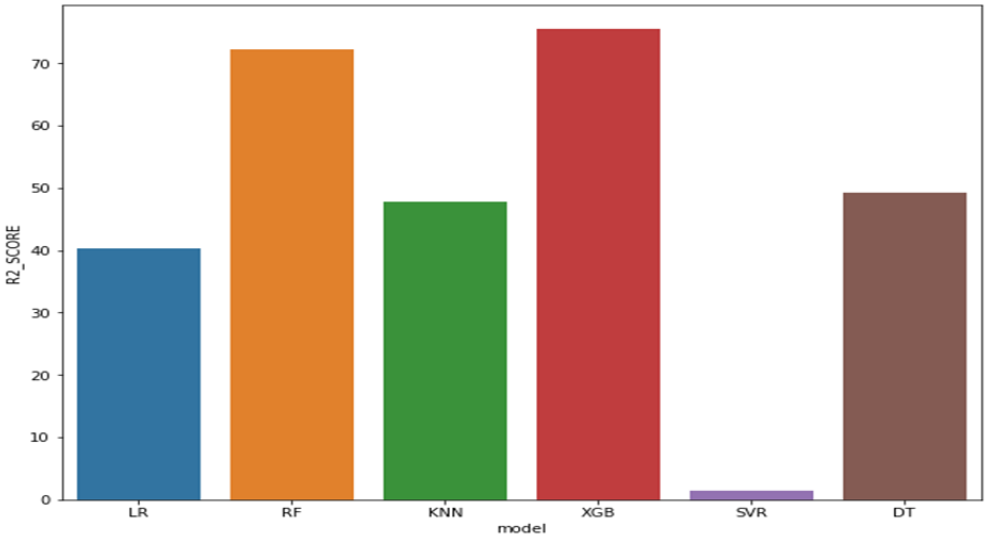

Figure 8 : Machine learning algorithms performance without pre-processing

From the table 3 it can be observed that random_state=0 will provide good performance for all algorithm with high R2 value and less RMSE value.

1.6 Feature Scaling on Air Quality Data of Gandhinagar

Four different feature scaling approaches named (1) Normalizer, (2) Robust Scaler, (3) Min-Max Scaler, and (4) Standard Scaler were applied. Figure 8 shows the effect of scaling on the Air Quality Data of Gandhinagar City and it is observed that robust scaler and standard scaler scale down to the near to same scale.

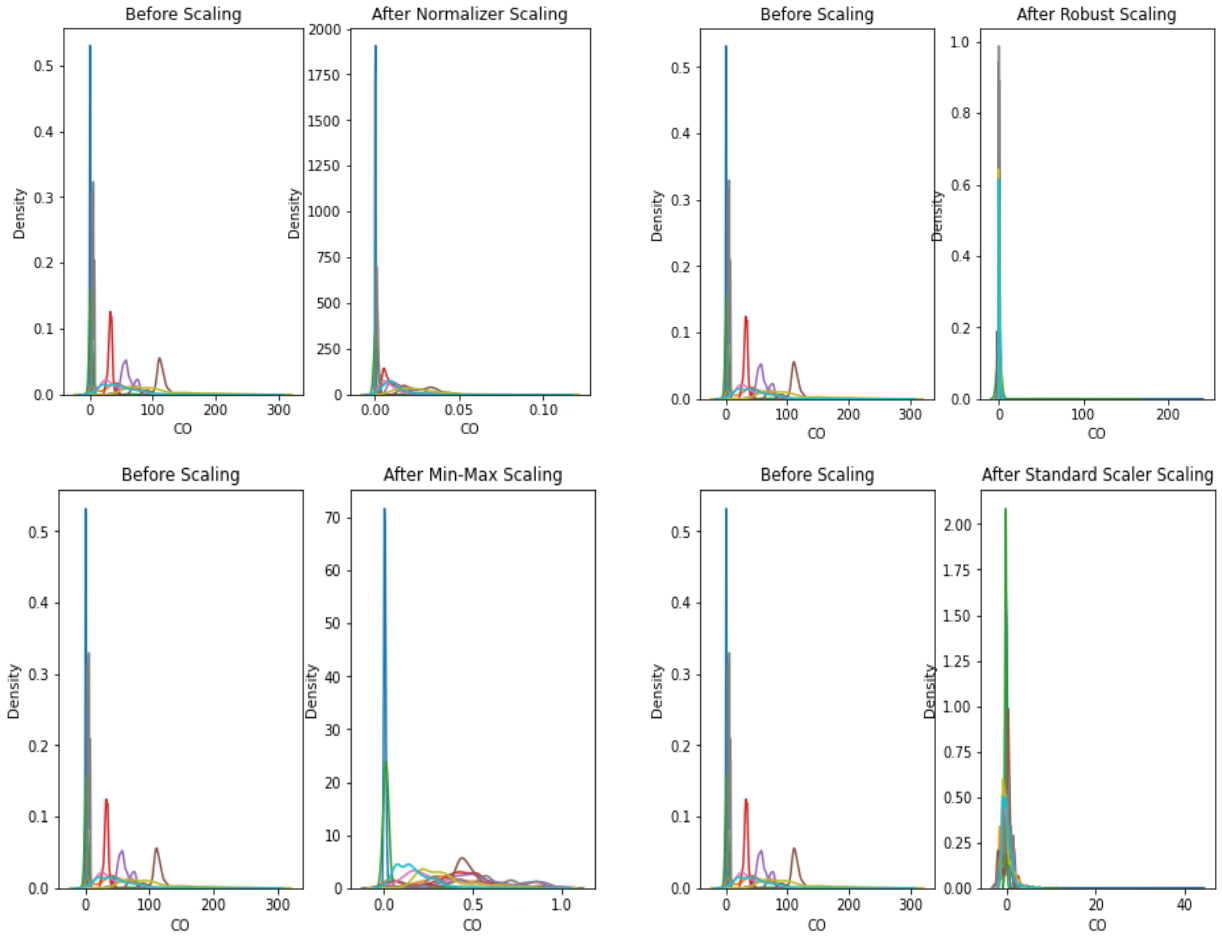

Figure 9 : Effect of scaling on Air Quality data of Gandhinagar

1.7 Machine Learning Algorithm performance after Feature Scaling

PM10 Model	Median Imputation		Normalizer		Robust Scaling		Min-Max Scaling		Standard-Scaler	
	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
LR	40.30	33.17	30.17	35.51	40.30	33.17	40.30	33.17	40.30	33.17
RF	71.69	22.84	62.14	25.45	71.74	22.82	71.83	22.78	71.55	22.90
KNN	50.31	30.26	56.58	27.60	53.85	29.17	46.22	31.49	52.15	29.70
XGB	73.88	21.94	62.54	27.23	73.88	21.94	73.87	21.95	73.88	21.94
SVR	0.01	42.93	2.37	42.74	27.00	36.68	32.98	35.15	37.87	33.84
DT	42.68	32.50	35.87	36.30	42.59	32.53	42.50	32.55	42.73	32.49

Table 5 : Machine learning algorithm performance with four feature scaling techniques.

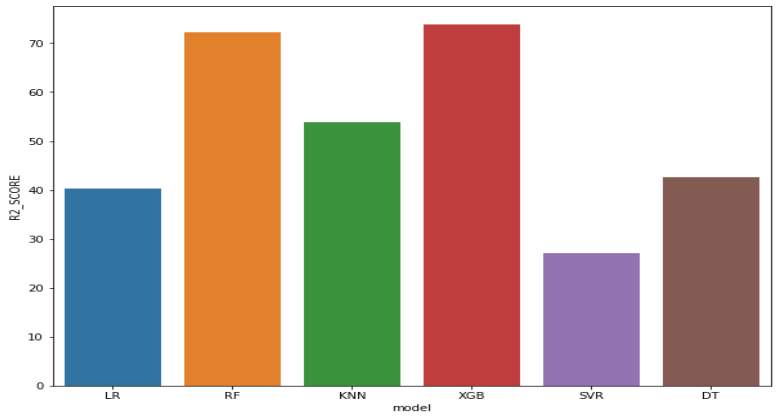

Figure 10 : Machine learning algorithms performance after applying feature scaling.

From Table 5, it was observed that Random Forest, XGBoost, Decision Tree, and Multiple Linear models are not much benefited from Feature scaling approaches while KNN and SVR benefited from Feature Scaling approaches. Out of the four Feature scaling methods Robust scaling gives good performance because it is not sensitive to outliers.

IV. CONCLUSION

In this study, feature scaling techniques, (1) Normalizer, (2) Robust Scaler, (3) Min-Max Scaler, and (4) Standard Scaler were applied to Gandhinagar air quality dataset. Six machine learning algorithms, Linear regression, Random Forest, K- Nearest Neighbor, XG-boost, Support Vector Regressor and Decision Tree Regressor were trained on dataset with and without feature scaling to compare performance for Particulate Matter Prediction. From analysis of Gandhinagar air quality dataset, it was observed that data are not normally distributed and dataset have outliers present. It is also identified that tree-based algorithms not get much benefits from feature scaling. Robust scaling gives good performance out of four feature scaling techniques on Gandhinagar air quality dataset. For PM10 prediction, without applying Feature scaling the XGBoost regressor with random state=0 model gives good accuracy with R2 score 74.58 and RMSE 21.92. Prediction model build with four feature scaling techniques in combination of six machine learning algorithms. For PM10 prediction after applying feature scaling, XGBoost algorithm in combination of Robust scaling gives good prediction accuracy with R2 score 73.88 and RMSE 21.94. so, it is concluded that it is not necessary that every dataset get benefit of feature scaling to improve machine learning algorithm performance.

Acknowledgments: The authors gratefully acknowledge the Gandhinagar Smart City Development Limited (GSCDL), Gandhinagar, Gujarat, India for the air quality data.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LR	Linear Regression
ML	Machine Learning
GSCDL	Gandhinagar Smart City Development Limited
PM	Particulate Matter
KNN	K nearest Neighbour
DT	Decision Tree
RF	Random Forest
XGB	XGBoost
SVR	Support Vector regressor
RMSE	Root Mean Square Error
R2	Coefficient of determination

V. REFERENCES

[1]. M. Mahajan, S. Kumar, B. Pant, U. K. Tiwari and R. Khan, "Feature Selection and Analysis in Air Quality Data," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 280-285, doi: 10.1109/Confluence51648.2021.9376882.

[2]. Particulate Matter (PM) Basics | US EPA. (2016, April 19). US EPA. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM

[3]. Gokhale S, Raokhande N (2008) Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period. Sci Total Environ 9–24

[4]. Report on the Environment, "Particulate Matter Emissions"- United states Environment protection Agency https://cfpub.epa.gov/roe/indicator_pdf.cfm?i=19

[5]. J Brownlee -," Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python",2020

[6]. Djarum, D.H., Ahmad, Z., Zhang, J. (2021). Comparing Different Pre-processing Techniques and Machine Learning Models to Predict PM10 and PM2.5 Concentration in Malaysia. In: Zaini, M.A.A., Jusoh, M., Othman, N. (eds) Proceedings of the 3rd International Conference on Separation Technology. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-16-0742-4_25

[7]. Y. Rybarczyk and R. Zalakeviciute, "Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review," Applied Sciences, vol. 8, no. 12, p. 2570, Dec. 2018.

Cite this article as :

Zalak L. Thakker, Dr. Sanjay H. Buch, "Effect of Feature Scaling Pre-processing Techniques on Machine Learning Algorithms to Predict Particulate Matter Concentration for Gandhinagar, Gujarat, India", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 11 Issue 1, pp. 410-419, January-February 2024. Available at doi : https://doi.org/10.32628/IJSRST52411150

Journal URL : https://ijsrst.com/IJSRST52411150