International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8070
***Article Received:** 23 August 2023 **Revised:** 31 August 2023 **Accepted:** 04 September 2023*

Implementation of Feature Engineering in Prediction of AQI in India using Machine Learning

Reema Gupta[1],Dr. Priti Singla[2]
[1]Research Scholar,Department of Computer Science and Engineering
[2]Faculty of Engineering, Department of Computer Science and Engineering
[1,2]Baba Mastnath University,Rohtak,India
reema2405@gmail.com

Abstract—Prediction of Air Quality Index (AQI) is the necessity of today's era but for the prediction, analysis of different preprocessing techniques that can be applied, needs to be considered. In this study, first of all we explored various feature engineering techniques such as Data Imputation, Scaling, Extraction, Selection, and Data Split that can be used before applying machine learning algorithm for better results. Second, we used MLR and SVR (Linear, Gaussian) to build the prediction models. Finally, we used root mean square error (RMSE), R^2, Mean Squared Error (MSE) and Mean Absolute Error (MAE) to evaluate the performance of the regression models in collaboration with the feature engineering techniques. The results shows that the performance of Linear SVR is better when coupled with imputation and robust scaler (R^2=0.7557834846394744) as compared to the others, the performance of Gaussian SVR is better when coupled with the imputation only as compared to the others. In case of MLR, results (R^2=0.7769187383819041) are almost same in all the 4 cases and performance degraded when PCA was applied.

Keywords-Imputation,AQI,Standard Scaler,Modeling,Evaluation.

I. INTRODUCTION

Air Pollution Index (AQI) was created to make it easier for people to comprehend the impact of local air quality on their health. It is a health protection alarming tool made to assist people in making decisions about how to safeguard their health by changing their activity levels and reducing their short-term exposure when air pollution levels are high. Jind is one of the most polluted cities in the state of Haryana, India. As per the Information provided by IQAir, India was in the top 5 polluted country among 118 countries in 2021.

Traffic related pollution is one of the major sources of air pollution. There is correlation between air quality and traffic CO intensity because pollution increases in the traffic peak hours i.e. in the morning and the evening [1]. In case of travelling from one place to the other, Air quality varies location to location. IOT based technology [2] can predict the air quality of the entire route and the destination place. Dust that is created during the demolition of old buildings and the following construction of new ones is also one of the main reasons of deteriorating the air quality.

Air quality prediction for a particular location can be done with the help of image capturing through daily used devices [3]. There is a need to look towards the harmful pollutants that affects the respiratory system and causes breathing problem and other respiratory disease. [4]Air is the basic need of every organism because polluted inhalation creates several health problems. Currently, laws are only put in place by the government when air quality reaches dangerous levels. If it is possible to predict when the air quality will reach dangerous levels, the Government can enact these rules quickly, possibly halting further deterioration of the air quality. This research tries to create a model that can review historical air quality data and predicts air quality index and amounts of various pollutants.

This paper is divided into sections. Section II includes the Literature Review, Section III discusses the Materials and Methodology, Section IV contains Results that displays the outcomes obtained after applying techniques and Section V is conclusion followed by the references.

II. LITERATURE REVIEW

For differentiating between the seasonal Air Quality, principal components analysis (PCA) was performed along-with construction of Ensemble models [5]**.** The study emphasized on identifying the air pollution sources in a span of five years in the city of Lucknow, India. The major source of pollution found were fuel combustion and emissions emitted from vehicles. Another research used Principal component regression (PCR) technique for forecasting daily AQI in Delhi with the help of previous day AQI and meteorological variables in four seasons in the years 2000-2006 [9]**.** They performed many statistical parameters which produced the same result but the performance of the PCR model was way better in the winter season as compared to any other season. They used the covariance of the input data matrix as a principal component.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8070
***Article Received:** 23 August 2023 **Revised:** 31 August 2023 **Accepted:** 04 September 2023*

An alternative method was proposed to the traditional sensor network. They suggest Machine learning and the Internet of Things (IoT) be applicable that use cloud-centric IoT middleware. It receives data from both- air pollution sensors as well as weather sensors. It is cost-efficient and more reliable. To monitor and predict air pollution, Artificial Neural Network (ANN) results in a reliable and suitable candidature [6]. Artificial neural network (ANN) used to predict air quality index (AQI) and air quality health index (AQHI) in span of one year in Ahvaz, Iran. Predictions were made based on hourly criteria of air pollutant concentrations and [10] concluded that ANN is reliable and can be used by practitioners to estimate air quality indices and spatial-temporal profile of pollutants.

An optimal hybrid model was proposed for forecasting of AQI by combining AI method and secondary decomposition (SD), optimization algorithm [11]. Their proposed idea successfully solved problems like considering influential factors based on decomposition technology. For their case study, they took two daily AQI series from Beijing and Guilin, China from December 2016- December 2018 and comprehended that their optimal-hybrid model has success-rate of forecasting AQI.

There are multiple issues discussed by [12] in predictions of AQI and accordingly the future needs to face such challenges. They also compared it with current research work on AQI which uses various models like machine learning and big data analysis. LightGBM model was proposed to predict the PM 2.5 concentration on the basis of 35 air quality stations that are monitoring in Beijing for 24 hours [13]. They compared the predicted data and the actual data. For their data source, they used the forecasting data. By using the lightGBM model they also resolved problems such as processing large scale and high-dimensional data. Their proposed model resulted as better alternative. The integration of predicted data improves stability of model and understands data adequately.

In a study, support vector regression (SVR) and random forest regression (RFR) models which are based on machine learning algorithms were used to predict the AQI of Beijing and the Italian city [8]**.** The results suggested that the RFR model is more reliable and time-efficient to perform complex and large samples. They established various models to improve the accuracy of the prediction of air indicators. Support vector regression (SVR) with Radial basis function (RBF) kernel was used to predict the concentration of various pollutants, ground level ozone and Air quality Index using already available data at US embassy and central pollution control board in New Delhi [14]. They tested various alternatives and found that radial basis function (RBF) helps in predicting hourly concentrations most accurately in the state of California [15].

Four different machine learning models namely Artificial neural network, statistical multilevel regression, deep learning long-short-term memory and neuro-fuzzy were applied on meteorological parametric dataset of 5 years [16]**.** The work concluded that Dl-LSTM is highly correlated with the dataset with low error levels and thus best suited. To increase the quality prediction performance, [7] suggested a deep learning algorithm. They used pre-processed data to find AQI. They demonstrated the utilization of data mining methods in natural resource research and environmental monitoring. Principal component analysis was applied to the deep learning models namely recurrent neural network (RNN), long short-term memory (LSTM) and bidirectional LSTM to forecast fine particulate matter (PM 2.5) in eight Korean cities for 5 years. The study [17] resulted in the conclusion that models with application of PCA produced better results. The PCA applied model is accurate for improving the performance of the model.

Numerous regression techniques and machine learning fused with big data analytics and IoT for the prediction of AQI namely SDG regression, Support vector regression, gradient boosting, linear regression, decision tree regression, adaptive boosting, random forest regression and artificial neural networks. Major air pollutants were used as a source to analyze the techniques. In the results [18] SVR and neural networks were found as best suited techniques.

III. MATERIALS AND METHOD

The flowchart of complete step by step process for deriving the output from given input data is presented in Fig1

A. RAW DATA

Dataset for the prediction of air quality is taken from the publicly available website of Central Pollution Control Board. The dataset used has 12 attributes and 851 instances of the Jind city in Haryana considered from 1st March, 2020 to 29th June, 2022. Each instance consist of the concentration of various pollutants, Ozone and other parameters such as Temperature, Wind Speed, Wind Direction, Relative humidity, etc. that are required as an input parameters and AQI as the target variable that helps to analyze the air quality.

B. DATA IMPUTATION

This step deals with the irregularities within the dataset, for example missing values or data having values ‘None’ or ‘NA’ are replaced with the mean of the corresponding column. Dataset is represented as D1.

```
D1 = D1.replace(to_replace='NA',value=np.nan)

D1 = D1.replace(to_replace='None ',value=np.nan)

D1['PM2.5'] = D1['PM2.5'].astype('float64')
```

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8070
***Article Received:** 23 August 2023 **Revised:** 31 August 2023 **Accepted:** 04 September 2023*

D1['PM2.5'].replace({np.nan: D1['PM2.5'].mean()},inplace = True)

This task is carried out on all components i.e. on various pollutants, Ozone and Meteorological parameters i.e. RH, WS, WD and AQI that are used in the prediction of the air quality index.

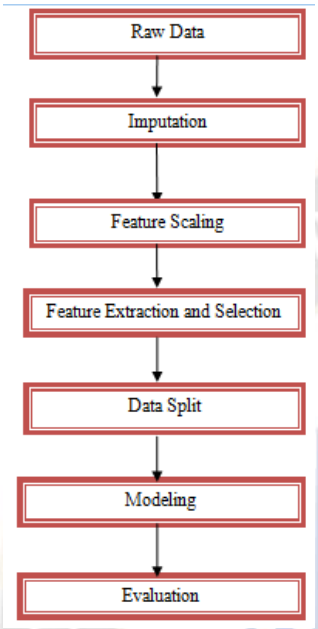

Fig. 1 Steps for Processing

C. FEATURE SCALING

This step standardizes the input variables in a fixed range and is a part of data preprocessing**.** For this, multiple techniques such as StandardScaler, Min max Scaler and Robust Scaler will be discussed in this paper.

1. **Standard Scaler-** This scaler transforms the each value in range of mean 0 and deviation 1.
2. **Minmax Scaler-** This scaler rescales each and every value to [0,1]
3. **Robust Scaler-** This scalers deals with handling of outliers

D. FEATURE EXTRACTION AND SELECTION

Feature Extraction is the process of reducing the dimensions of the dataset. It reduces the complexity and creates new ones which is linear combination of the existing ones.

Principal Component Analysis is one of the techniques that can be used for this purpose. PCA technique uses Covariance matrix, Eigen vectors and Eigen values.

Feature Selection is a method in which we have to select the best variables as input that can be used to build a prediction model. Output of the model is dependent on the quality of the data that are given as input to the model. Appropriate selection of the variables is the preprocessing step before applying the machine learning model.

E. DATA SPLIT

In this step given data set is splitted into train-test ratio. Splitting ratio should not undergo the over-fitting and under-fitting problems. In this study, train-test ratio considered is 80:20 respectively.

F. MODELING AND EVALUATION

For Modeling Machine learning algorithm, either classification or regression is applied on the problem and result obtained is compared with the actual value. If the difference between the actual value and the result is very low, this implies applied algorithm is working well in the environment. The performance of machine learning algorithms are evaluated through various metrics like accuracy, precision, Mean square error, Coefficient Metrics, etc.

The main objective is to predict the air quality index using machine learning technique. To create a prediction model two regression techniques are used i.e. MLR (Multiple linear regression) and SVR (Support Vector regression).

MLR (Multiple Linear Regression) – It is simply a regression algorithm in which the response variable is calculated based on multiple variables which serves as the input. Prediction of the response variable depends on how they are correlated with the independent variables.

SVR (Support Vector Regression) – SVR works on the same principle of SVM (Support Vector Machine) . The only difference is that, SVM deals with classification and regression problems while SVR deals with the regression problems.

IV. RESULTS

Python tool was used to obtain the results using machine learning algorithms. Numpy, pandas, Matplotlib, Scikit-learn, seaborn are the major libraries used in the implementation. Table I-III showing the performance of Linear SVR, Gaussian SVR and MultiLinear Regression algorithms are presented below wherein M1 represents Mean Absolute Error, M2 represents Mean Squared Error, M3 represents Root Mean Square Error and M4 represents Coefficient R^2. Different preprocessing techniques such as StandardScaler, Minmax Scaler, RobustScaler and PCA along with Imputation were applied for air quality prediction. Selection extracted the required components in the prediction from the dataset followed by the Data split in train-test ratio 80:20 respectively.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8070
***Article Received:** 23 August 2023 **Revised:** 31 August 2023 **Accepted:** 04 September 2023*

TABLE I
Performance Evaluation of Linear SVR while applying different Preprocessing techniques

Linear SVR				
Preprocessing Techniques	**M1**	**M2**	**M3**	**M4**
Imputation	33.04144965797549	2697.8424632036053	51.940759170458854	0.7330234447729286
Imputation and Standard Scaler	33.87663968082457	2472.5937100001697	49.72518184984515	0.7553138998382833
Imputation and Minmax Scaler	66.65289061446242	6406.411229474495	80.04006015411592	0.36602613950181295
Imputation and Robust Scaler	34.058417897389525	2467.848477537152	49.67744435392336	0.7557834846394744
Imputation and PCA	36.925125504159986	3087.234773408241	55.5628902542717	0.6944894610328636

TABLE II
Performance Evaluation of Gaussian SVR while applying different Preprocessing techniques

Gaussian SVR				
Preprocessing Techniques	**M1**	**M2**	**M3**	**M4**
Imputation	53.61567208810835	4741.171164960525	68.85616286840651	0.5308171019519038
Imputation and Standard Scaler	62.94170943020042	6346.393504078221	79.66425487053915	0.37196545056144925
Imputation and Minmax Scaler	65.13933618007908	6527.617399185145	80.79367177684861	0.3540316638780616
Imputation and Robust Scaler	58.46840816041663	5630.9255829013555	75.03949348777186	0.44276764289719117
Imputation and PCA	54.06578733773182	5237.580499642561	72.37113029131548	0.4816927902237098

TABLE III
Performance Evaluation of MLR while applying different Preprocessing techniques

MLR				
Preprocessing Techniques	**M1**	**M2**	**M3**	**M4**
Imputation	35.58786771686932	2254.2732257012267	47.47918728981391	0.7769187383819041
Imputation and Standard Scaler	35.587867716869305	2254.273225701229	47.479187289813936	0.7769187383819038
Imputation and Minmax Scaler	35.5878677168693	2254.2732257012285	47.47918728981393	0.7769187383819038
Imputation and Robust Scaler	35.5878677168693	2254.2732257012294	47.479187289813936	0.7769187383819038
Imputation and PCA	40.19107456509701	2785.609554106226	52.7788741269291	0.7243380764050298

A. Linear SVR Results

Fig 2-6 shows the actual and predicted AQI Values when using Linear SVR model and using imputation, imputation with scaling techniques and imputation with PCA as preprocessing technique to enhance the dataset before processing. The values shown in the below figures are of test dataset which includes actual values and predicted ones using the approach.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8070
Article Received: 23 August 2023 *Revised:* 31 August 2023 *Accepted:* 04 September 2023

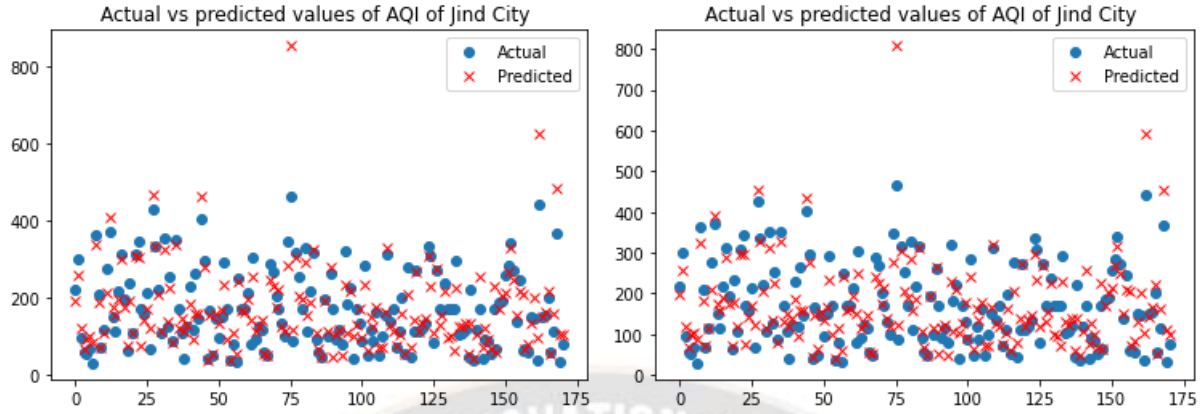

Fig 2 and 3 shows the actual and predicted values of AQI a) Imputation b) Imputation and Standard Scaler

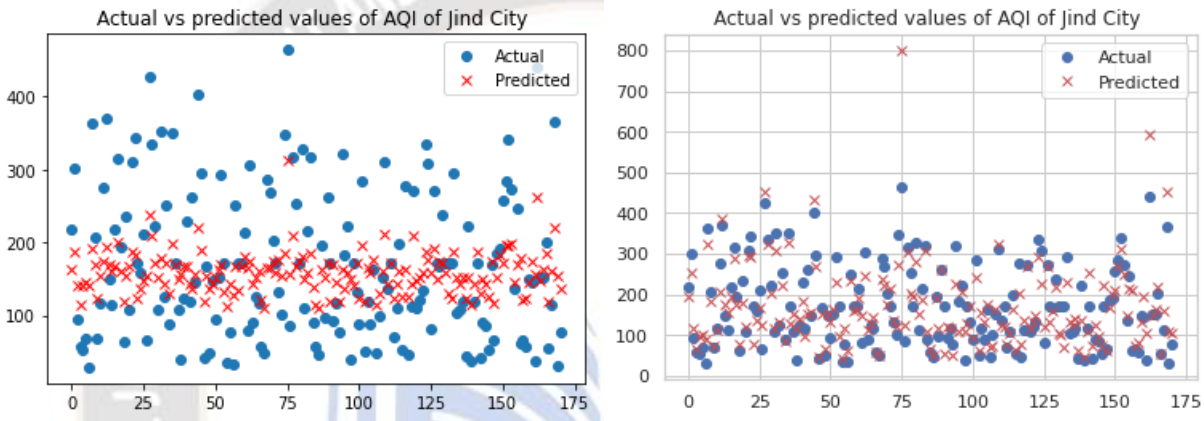

Fig 4 and 5 shows the actual and predicted values of AQI a) Imputation and minmax scaler b) Imputation and Robust Scaler

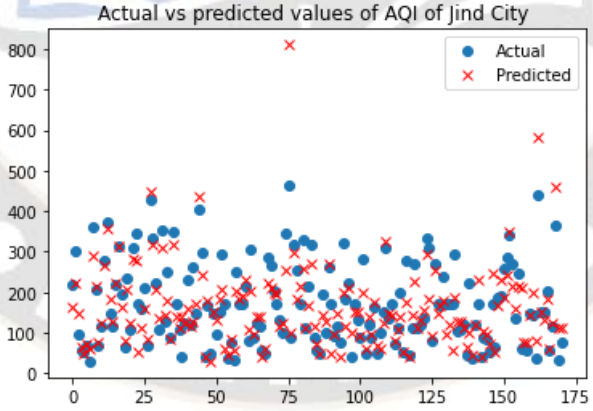

Fig 6 shows the actual and predicted values of AQI (Imputation and PCA)

B. Gaussian SVR Results

Fig 7-11 shows the actual and predicted AQI Values when using Gaussian SVR model and using imputation, imputation with scaling techniques and imputation with PCA as preprocessing technique to enhance the dataset before processing. The values shown in figures are of test dataset which includes actual values and predicted ones using different approaches.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8070
***Article Received:** 23 August 2023 **Revised:** 31 August 2023 **Accepted:** 04 September 2023*

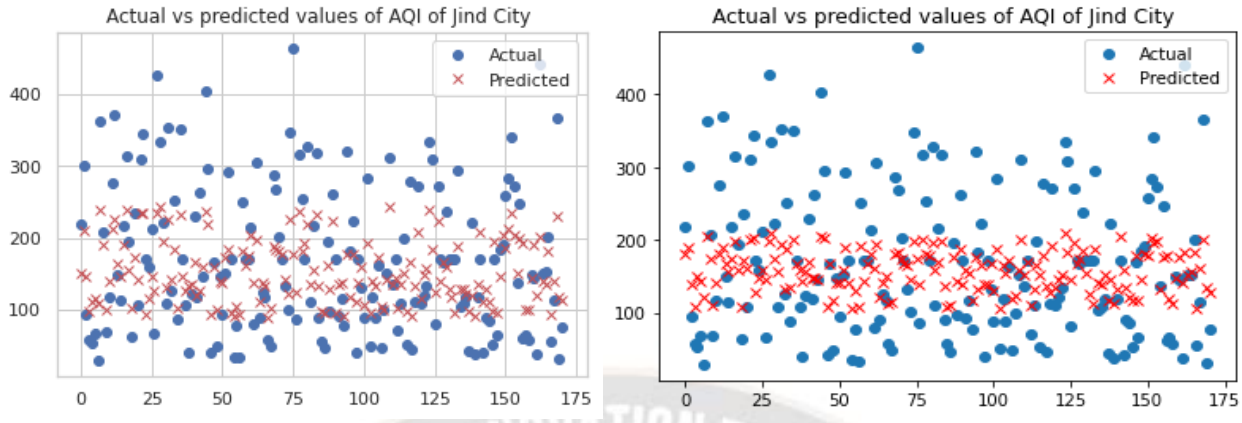

Fig 7 and 8 shows the actual and predicted values of AQI a) Imputation b) Imputation and Standard Scaler

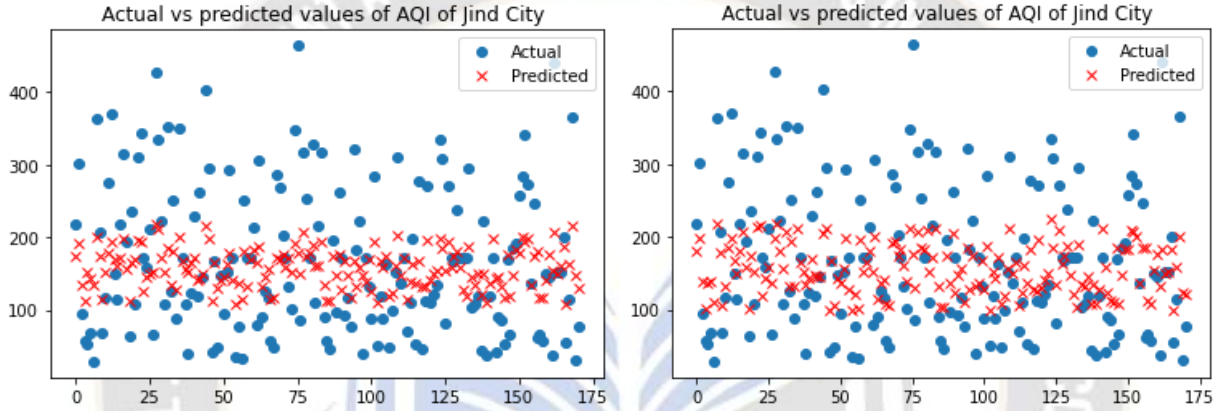

Fig 9 and 10 shows the actual and predicted values of AQI a) Imputation and minmax scaler b) Imputation and Robust Scaler

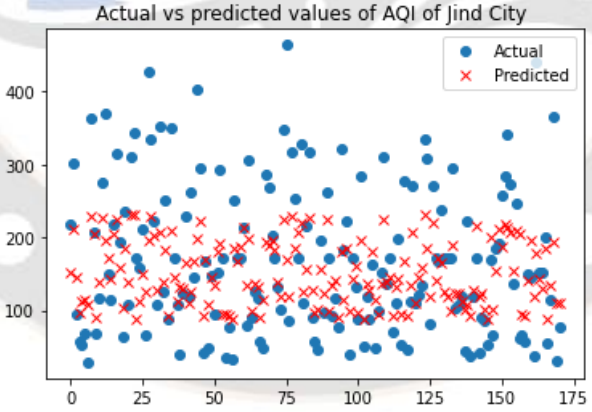

Fig 11 shows the actual and predicted values of AQI (Imputation and PCA)

C. MLR Results

Fig 12-16 shows the actual and predicted AQI Values when using MLR model and using imputation, imputation with scaling techniques and imputation with PCA as preprocessing technique to enhance the dataset before processing.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8070
***Article Received:** 23 August 2023 **Revised:** 31 August 2023 **Accepted:** 04 September 2023*

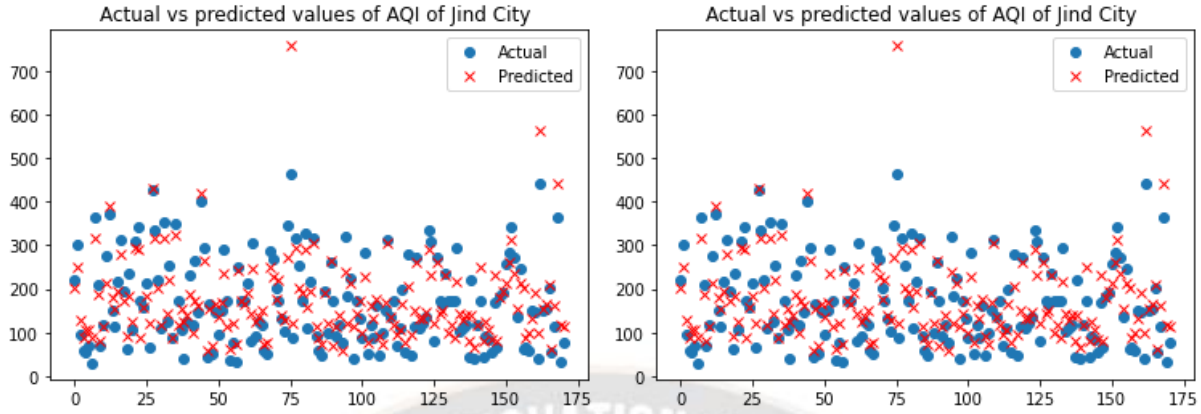

Fig 12 and 13 shows the actual and predicted values of AQI a) Imputation b) Imputation and Standard Scaler

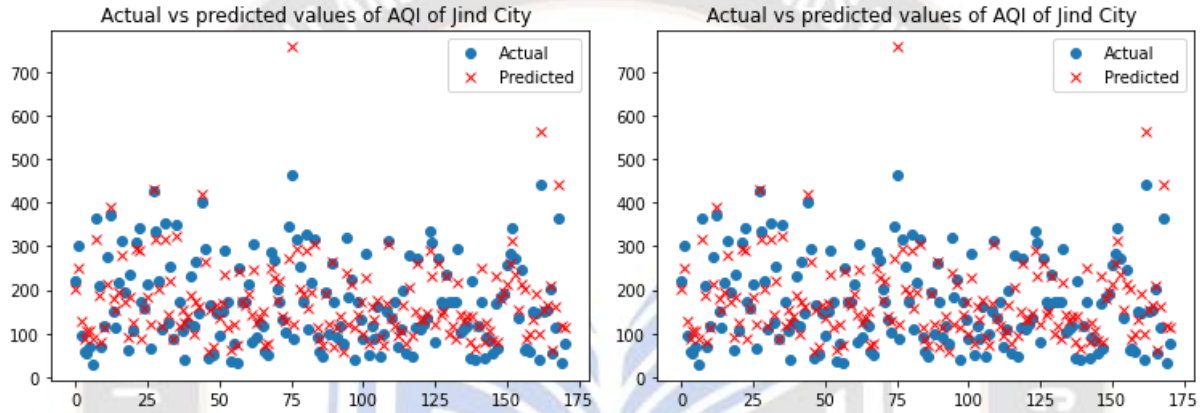

Fig 14 and 15 shows the actual and predicted values of AQI a) Imputation and minmax scaler b) Imputation and Robust Scaler

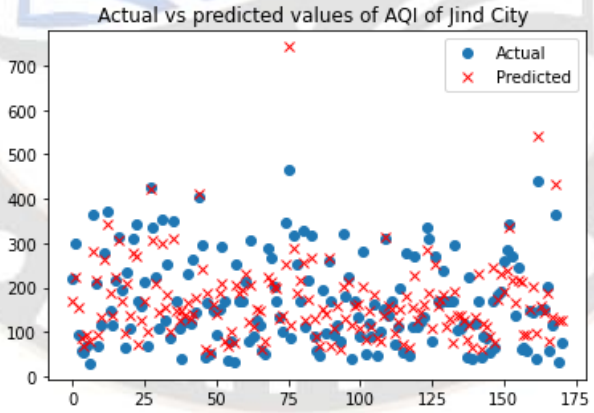

Fig 16 shows the actual and predicted values of AQI (Imputation and PCA)

V. COMPARATIVE ANALYSIS

Imputation with Robustscaler technique with MAE 34.058417897389525, RMSE 49.67744435392336, R^2 0.7557834846394744 outperforms as compared to other preprocessing technique with Linear SVR. In case of Gaussian SVR performance is best when only imputation technique with MAE 53.61567208810835, Coefficient R^2 0.5308171019519038 is used. In case of MLR, performance is same when the imputation technique is applied and when imputation is used along with the scaling techniques with MAE, RMSE and R^2 35.58786771686932, 47.47918728981391, 0.7769187383819041 respectively. These best approaches of each regression technique are compared based on evaluation metrics is shown in Fig 17.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8070
***Article Received:** 23 August 2023 **Revised:** 31 August 2023 **Accepted:** 04 September 2023*

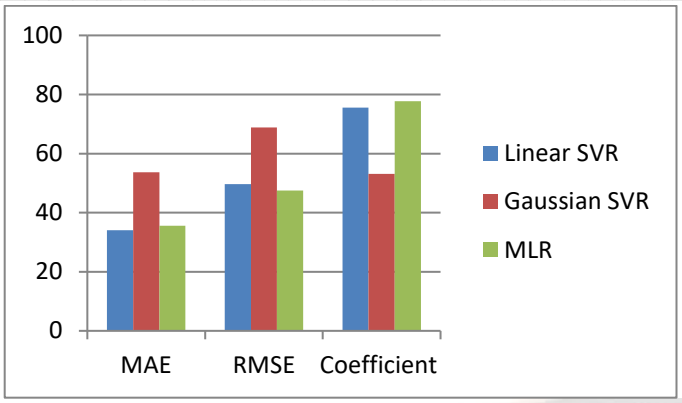

Fig 17 Comparison based on Evaluation Metrics

VI. CONCLUSION

Air quality index is predicted using Prediction models based on MLR, SVR. Dataset contained null values and outliers that need to be handled before applying model for the prediction. This paper concludes that the performance of the algorithm improves if outliers and irregular values in the dataset can be handled using above discussed techniques. Transformation of raw data into the fruitful data is one of the needs of the machine learning model. In this study PCA degraded the performance of the existing model in the prediction. RMSE, R^2, MSE and MAE have been used to evaluate the performance of the regression models when different preprocessing techniques were applied to improve the performance. In case of Linear SVR, imputation and robustscaler when applied gave better result as compared to the others while results produced using Gaussian SVR is better when applied imputation only as compared to the others. In case of MLR, performance results are almost same in all the 4 cases and performance degraded when PCA was applied. Overall the performance of MLR is best as compared to the others in terms of various evaluation metrics used in the paper.

REFERENCES

[1] L. Pan, E. Yao and Y. Yang, "Impact Analysis of Traffic-Related Air Pollution based on Real time Traffic and Basic Meteorological Information," Journal of Environmental Management, vol. 183, no. 3, pp. 510-520, 2016.

[2] S. Dhingra, R. B. Madda, A. H. Gandomi, R. Patan and M. Daneshmand, "Internet of Things Mobile-Air Pollution Monitoring System (IoT-Mobair)," IEEE Internet of Things Journal, vol. XX, no. XX, p. 8, 2019.

[3] J. Ma, K. Li, Y. Han and J. Yang, "Image based Air Pollution Estimation Using Hybrid Convolutional Neural Network," in 24th International Conference on Pattern Recognition, Beijing, China, 2018.

[4] R. Brugha and J. Grigg, "Urban Air Pollution and Respiratory Infections," Paediatric Respiratory Reviews, vol. 15, no. 2, p. 6, 2014.

[5] K. P. Singh, S. Gupta and P. Rai, "Indentifying Pollution Sources and Predicting Urban Air Quality using Ensemble Learning Methods," Atmospheric Environment, vol. 80, pp. 426-437, 2013.

[6] I. U. Samee and M. T. Jilani, "An Application of IOT and Machine Learning to Air Pollution Monitoring in Smart Cities," IEEE, p. 6, 2019.

[7] S. K, A. K, G. M, G. R and M. A, "Air Quality Prediction Using Classification Techniques," Annals of R.S.C.B, vol. 25, no. 4, pp. 3794-3805, 2021.

[8] H. Liu, Q. Li, D. Yu and Y. Gu, "Air Quality Index and Air Pollutant Concentration Prediction based on Machine Learning Algorithms," MDPI, no. 4069, p. 9, 2019.

[9] A. Kumar and P. Goyal, "Forecasting of air quality in Delhi using principal component regression technique," Atmospheric Pollution Reasearch 2, pp. 436-444, 2011.

[10] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. T. Birgani and M. Rahmati, "Air Pollution Prediction using an Artificial Neural Network Model," Clean Technologies and Environmental Policy, vol. 21, pp. 1341-1352, 2019.

[11] Q. Wu and H. Lin, "A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors," Science of the total environment, vol. 683, pp. 808-821, 2019.

[12] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu and G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," International Journal of Environmental Science and Developement, vol. 9, no. 1, pp. 8-16, 2018.

[13] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang and L. Huang, "A Predictive Data Feature Exploration - Based Air Quality Prediction Approach," IEEE Access, vol. 7, pp. 30732-30743, 2019.

[14] S. Bhattacharya and S. Shahnawaz, "Using Machine Learning to Predict Air Quality Index in New Delhi," arXiv:2112.05753, p. 7, 2021.

[15] M. Castelli, F. M. Clemente, A. Popovik, S. Silva and L. Vanneschi, "A Machine Learning Approach to Predict Air Quality in California," Wiley, vol. 2020, p. 23, 2020.

[16] C. AmuthaDevi, D. S. Vijayan and V. Ramachandran, "Development of Air Quality Mornitoring (AQM) Models using different Machine Leaning Approaches," Journal of Ambient Intelligence and Humanized Computing, p. 13, 2021.

[17] S. W. Choi and B. H. Kim, "Sustainability," Applying PCA to Deep Learing Forecasting Models for Predicting PM 2.5, vol. 13, no. 7, p. 30, 2021.

[18] C. Srivastava, S. Singh and A. P. Singh, "Estimation of Air Pollution in Delhi using Machine Learning Techniques," in 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, 2018.