International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Assessment of Air Pollutants of Dhanbad Using Machine Learning Techniques

**Uday Kumar Sinha[1, *], K. Bandyopadhayay[2] and S. C. Dutta[3]**

**Submitted**:07/11/2023 **Revised**: 11/12/2023 **Accepted**: 19/01/2024

**Abstract:** A comprehensive assessment of air pollution in Dhanbad for the months of April 2019 through March 2023 was conducted using a support vector machines. random forest, xgboost, and decision tree methods in relation to seven main pollutants into the air (PM10, NO, NO2, NH3, SO2, CO, and O3). A randomly selected 30-day period was used to create line and bar graphs using accuracy & error matrices for 7 pollutants in various models. Our investigation found that the Random Forest model estimates Dhanbad air contaminants with the lowest MAE.Among the aforementioned models, the Random Forest one stands head and shoulders above the others.

***Keywords:*** *Machine learning models, air quality, air pollutants, CPCB, data preprocessing, performance criteria*

## 1. Introduction

Air is one of the basic elements of ***PanchTatva*** for surviving on the surface of the earth because the life systems are dependent on smooth as well as harmonious functions of basic five elements of PanchTatva. It is therefore, necessary to maintain purity as well as quality of air for the existence of civilization. However, quality of air has been deteriorating continuously because of unplanned industrial as well as infrastructural development and deforestation in many countries including India. In particular, increase of $PM_{10,}$ CO, $SO_2$, NO, $NH_3$ etc. have attained threatened level and hence, reduction of the quantity of air pollutants are utmost urgent for the betterment of the society and civilization. It is therefore, necessary to analyze the air quality of different places and necessary measures should be taken accordingly.

In this work, we consider sevenair pollutants such as $PM_{10}$[1], NO[2]], $NO_2$[3], $NH_3$[4], $SO_2$[5], CO[6], and $O_3$[7]for comparison of different models of Machine Learning algorithm. We have chosen five important models like Linear Regression, Support Vector Machine, Random Forest, XGBoostand Decision Tree foranalyzing the data available for estimation of air pollutants of Dhanbad.

Data on the seven pollutants have been gathered throughout the period of time from the Central Pollution Control Board (CPCB)[8]website from April 2019 to March 2023 in order to analyze the air pollutants of Dhanbad.

According to the current investigation, the Random Forest algorithm is the most accurate model for forecasting pollutant concentrations, with the exception of NO2, where XGBoost demonstrates the highest accuracy. Thus, based on data available throughout the aforementioned annular period, it can be said without reservation that the Random Forest technique provides highest accuracy with minimal mistakes for measuring the air contaminants of Dhanbad.

The plan for the paper is presented after this. In the second section, 5 machine learning algorithms are briefly explained. The third section covers the work's methodology, while the fourth section covers results and a summary.A conclusion is in the fifth section.

## 2. Machine Learning Models

Machine learning models show how computers can categorise, anticipate, and evaluate physical occurrences and their qualities by training them with complex data. This research builds a variety of models using various methods to predict efficient pattern recognition & model self-learning. Five algorithms' fundamentals are presented here.

*1Research Scholar, University Department of Physics, BBMK University, Dhanbad, Jharkhand, 828103*
*&*
*Department of Vocational Studies, Guru Nanak College, Dhanbad, Jharkhand, 826001*
*ORCID ID: 0000-0003-2858-2463*
*2. Associate Professor, University Department of Physics, BBMK University, Dhanbad, Jharkhand, 828103*
*ORCID ID: 0009-0001-3356-1811*
*3. Associate Professor, Department of CSE and IT, BIT – Sindri, Dhanbad, Jharkhand, 828123*
*ORCID ID: 0009-0002-9947-3619*
**** Corresponding Author** Email: udaydhn@gmail.com*

### 2.1. Linear Regression *(LR)*

Linear regression is the first machine learning method most academics use [9]."Linear regression" is supervised machine learning that identifies the linear connection between a dependent variable & one or more variables that are independent. The approach finds the best linear equation to predict the dependent variable given the independent variables. The equation shows a straight line between independent and dependent variables.

### 2. 2. Support Vector Machine [SVM]

Support vector machines are supervised learning methods for regression, classification, and more [10].The hyperplane divides data points into intervals from which the model output may be calculated. SVMs come in two varieties. Regression problems are different from classification problems. Most issues are nonlinear, therefore the kernel approach converts low-dimensional data to high-dimensional data multidimensional feature space. Kernels are used to separate indivisible input data. such as linear, Use polynomial, RBF, sigmoid, hyperbolic tangent, etc. Radial Basis Function Kernel is more accurate than other kernels, hence we used it [11].

### 2.3. The Random Forest (RF)

An other well-known machine learning approach is random forest[12]. This method employs supervised ensemble learning to finish problems related to regression or classification. Using a training set of data, it constructs many decision trees. The mean value of the decision tree set is then used to forecast the value for fresh input data.

### 2.4. XGBoost, or extreme gradient boosting

Another significant machine learning method is XGBoost[13]. The machine learning technique XGBoost combines a set of decision trees with gradient boosting to generate predictions. The XGBoost method uses the community-based weak learning technique to solve regression and classification problems. This method produces useful results since it relies predictions on parallel tree structures and takes hardware and software parameters into account while building them.

### 2.5.Decision Tree algorithm or DT

Decision tree machine learning is used for regression and classification[14].This tree-like model has leaf nodes that represent class labels or numerical values, inner nodes that represent attributes or features, and branch nodes that provide decision rules depending on those features.

## 3. Methodology

Present section deals with the collection and Pre-processing of data, designing of models and criteria of performance for the current work.

### 3.1. Assemblage of Data

Datasets for the current investigation were gathered from the Central Pollution Control Board (CPCB), New Delhi, website as well asContinuous Ambient Air Quality Monitoring Station (CAAQMS),Dhanbad, Jharkhand. Particulate matter ($PM_{10}$), nitric oxide (NO), nitrogen dioxide ($NO_2$), ammonia ($NH_3$), sulphur dioxide ($SO_2$), carbon monoxide (CO), and ozone ($O_3$) are the seven features in the datasets, which also include meteorological parameters and pollutant concentrations.

### 3.2. Pre-processing of Data

High-quality data and effective data representations are key factors in a model's effectiveness.The following explanations pertain to crucial data pre-processing procedures, which include missing value incorporation, outlier elimination, feature scaling, and feature selection [15].

**(i) Incorporation of Missing Values:**The nearest data points should be taken into consideration for analysis in lieu of the missing data, which should be replaced by linear interpolation estimation.

**(ii) Removal of Outliers:** For CO, an unusual pattern was seen for the period from October to November in 2019. We have not considered values of CO for this period in the present work.

**(iii) Features scaling:** One crucial stage in the pre-processing of data before creating a model is feature scaling. The range of characteristics is normalised or standardised using it. Min-max scaling has been used in this research to normalise the values between 0 and 1.The following is the normalisation formula: $Z' = \frac{Z - Z_{min}}{Z_{max} - Z_{min}}$

Z is the initial value, Z^' is the normalised value, 〖Z〗_max is the maximum value that can be achieved, & Z_min is the lowest value that can be achieved for the feature.

**(iv)Selection of features:** A portion of the extensive feature space's features are selected in order to lower the dataset's dimensionality. This phase uses Pearson's Correlation coefficient to choose the characteristics.

### 3.3.Designing of Model

The model is created on a 64-bit machine running Python 3.7.6[16] on a 2.2GHz Intel Core i3 CPU with 4GB RAM. Using common charting tools like matplotlib[17],

seaborn[18], and sklearn[19], the model is trained and forecasted. A large portion of the work was implemented using Panda's dataframe. We were able to use the training set to train and evaluate the previously mentioned models by splitting the dataframe into a testing set and a training set with a ratio of 7:3.

### 3.4. Models'Performance Standards

Prediction models are examined for effectiveness using statistical matrices like Root Mean Squared Error or RMSE, Mean Absolute Error or MAE, Mean Squared Error or MSE, & R-Squared or $R^2$[20].

**R-Squared or $R^2$:** R-Squared, often known as $R^2$, is a statistical measure that illustrates how well the model and data fit in a regression model. A higher $R^2$ value indicates a better model. The formula used to determine $R^2$ is

$$R^2 = 1 - \frac{(x_i - \hat{x_i})^2}{(x_i - \bar{x_i})^2}$$

**Mean Absolute Error /MAE** is a statistical phrase that measures the average of the absolute differences between the observed values and the values predicted for the whole collection of data. The formula used to calculate the MAE is

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - \hat{x_i}|$$

**Mean Squared Error/MSE** total of the squared deviations for all values in the data set, both observed and anticipated.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{x_i})^2$$

**Squared Error of Root Mean:** As one of the most frequently employed metrics for assessing the precision of predictions, RMSE (root mean squared error) is frequently utilised. Sqrt (mean squared error) It may be expressed as follows:

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

The expected x-value is represented by by $x\perp\hat{}\_$, the actual value of x is represented by x_i in all of the above formulas, and the number of mistakes is indicated by 'n'.

### 4. Result and Summary

The accuracy, mean squared error, root mean square error, & mean absolute error are now shown for each of the five machine learning models, which cover seven distinct contaminants, that we previously examined. To make things simple, line and bargraphs covering a 30-day period selected at random have been made. The analytical assessment of the air quality in Dhanbad is shown in the diagrams.

### 4.1 $PM_{10}$

Estimation of accuracy and error metrics for $PM_{10}$ in different models

| Models | Coefficient of determination R^2 on Test Data |
|---|---|
| Linear Regression | 0.713524 |
| Decision Tree | 0.56395 |
| Random Forest | 0.738381 |
| Support Vector Machine | 0.657383 |
| XGBoost | 0.71769 |

**Table 1 (a):** Estimation of accuracy for $PM_{10}$ in different models

| Models | MAE | MSE | RMSE |
|---|---|---|---|
| Linear Regression | 0.0476522 | 0.00580864 | 0.0762145 |
| Decision Tree | 0.0570792 | 0.00904255 | 0.0950923 |
| Random Forest | 0.0445463 | 0.00529851 | 0.0727909 |
| Support Vector Machine | 0.0595149 | 0.00694697 | 0.0833485 |
| XGBoost | 0.0478898 | 0.00572418 | 0.0756583 |

**Table 1 (b):** Estimation of error metrics for $PM_{10}$ in different models

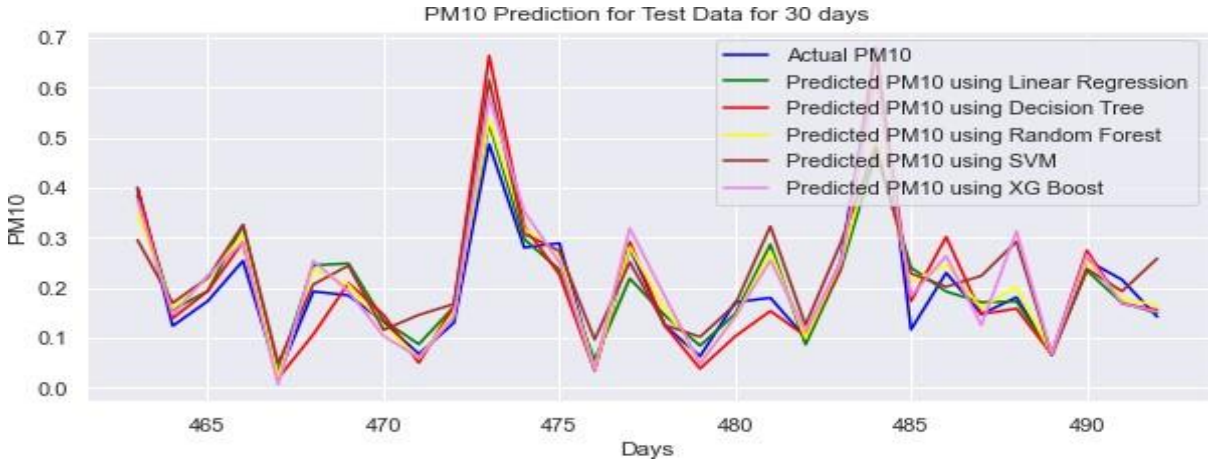


**Fig 1 (a):** Line graph Comparing $PM_{10}$levels between the models' actual and predicted levels

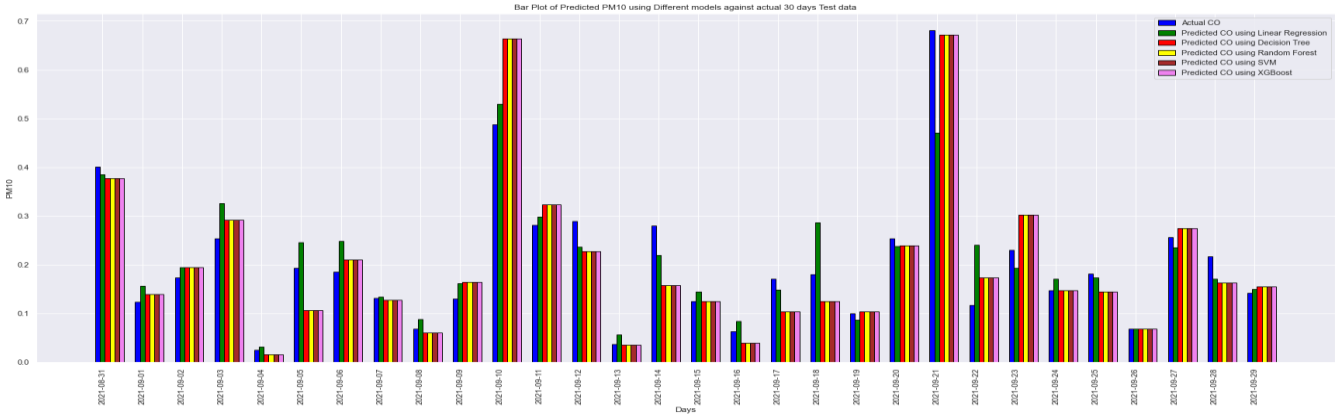


**Fig 1 (b):** Bar graph comparing $PM_{10}$ levels between models' actual and anticipated values

### 4.2. $NO_2$

Estimation of accuracy and error metrics for $NO_2$ in different models

| Models | Coefficient of determination R^2 on Test Data |
|---|---|
| Linear Regression | 0.886757 |
| Decision Tree | 0.940375 |
| Random Forest | 0.961305 |
| Support Vector Machine | 0.808075 |
| XGBoost | 0.966593 |

**Table 2 (a):** Estimation of accuracy for $NO_2$ in different models

| Models | MAE | MSE | RMSE |
|---|---|---|---|
| Linear Regression | 0.0205374 | 0.00186783 | 0.0432185 |
| Decision Tree | 0.00847022 | 0.000983462 | 0.0313602 |
| Random Forest | 0.00894517 | 0.000638241 | 0.0252634 |
| Support Vector Machine | 0.0465254 | 0.00316562 | 0.0562639 |
| XGBoost | 0.00917657 | 0.000551015 | 0.0234737 |

**Table 2 (b):** Estimation of error metrics for $NO_2$ in different models

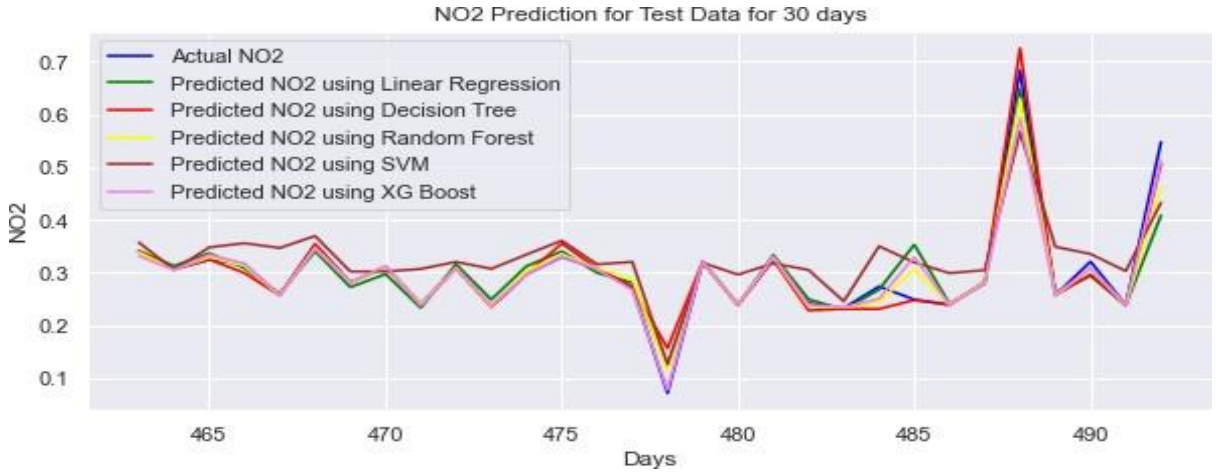


**Fig 2 (a):** A line graph that compares the models' real and anticipated $NO_2$ levels

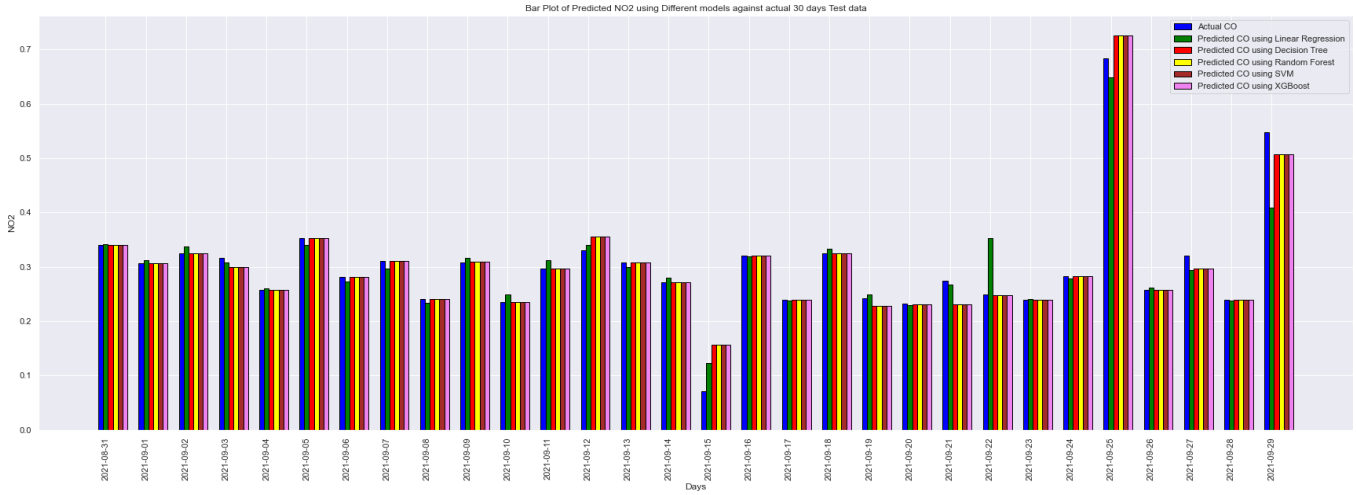


**Fig 2 (b):** Bar graph showing how the models' actual and anticipated $NO_2$ levels compare

### 4.3.NO

Estimation of accuracy and error metrics for NO in different models

| Models | Coefficient of determination R^2 on Test Data |
|---|---|
| Linear Regression | 0.846018 |
| Decision Tree | 0.746021 |
| Random Forest | 0.850919 |
| Support Vector Machine | 0.799639 |
| XGBoost | 0.822513 |

**Table 3 (a):** Estimation of accuracy for NO in different models

| Models | MAE | MSE | RMSE |
|---|---|---|---|
| Linear Regression | 0.0241867 | 0.00255732 | 0.0505699 |
| Decision Tree | 0.0269907 | 0.00375439 | 0.0612731 |
| Random Forest | 0.0222771 | 0.00254159 | 0.0504142 |
| Support Vector Machine | 0.0452136 | 0.00332758 | 0.0576851 |
| XGBoost | 0.0245447 | 0.00294768 | 0.0542926 |

**Table 3 (b):** Estimation of error metrics for NO in different models

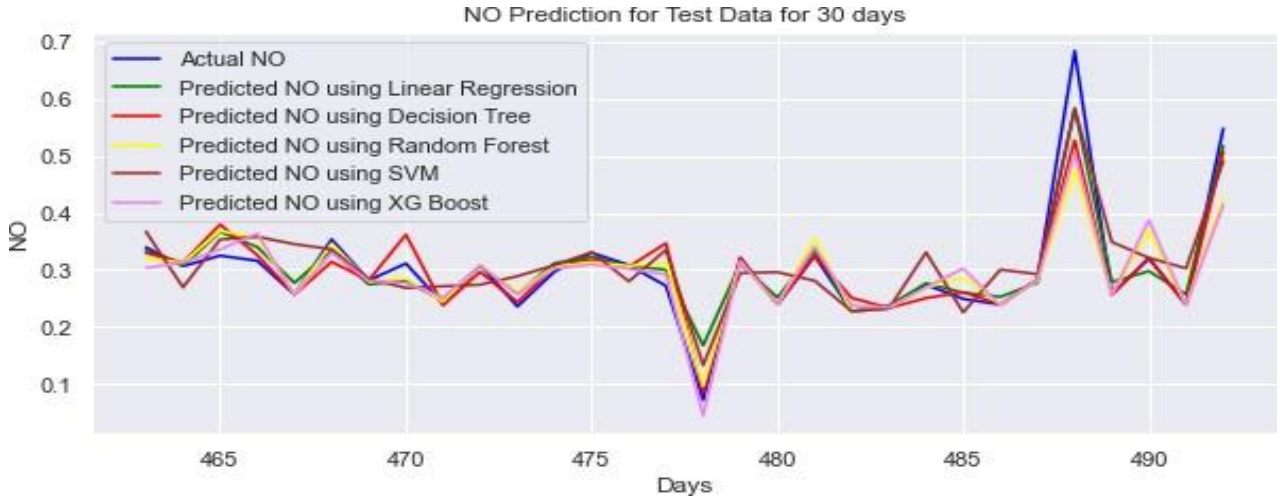


**Fig 3 (a):** A line graph comparing the models' actual and anticipated levels of NO

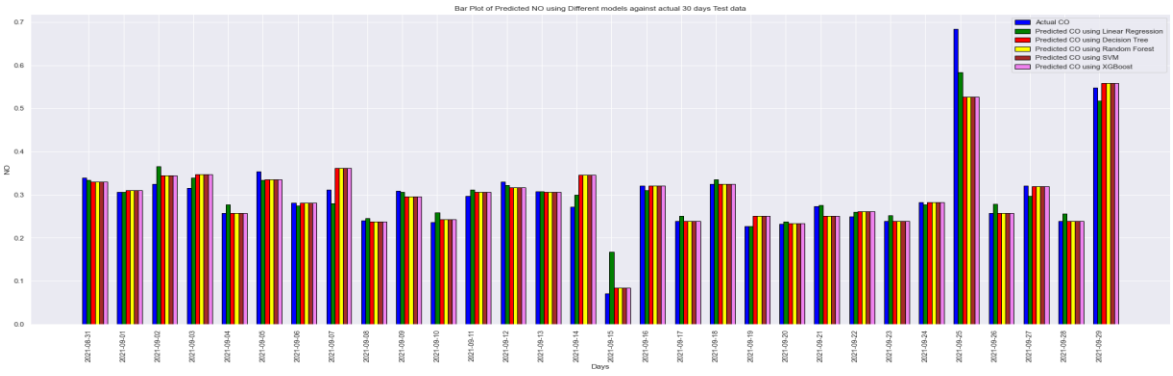


**Fig 3 (b):** Bar graph comparing the models' actual and anticipated levels of NO

### 4.4. $NH_3$

Estimation of accuracy and error metrics for $NH_3$ in different models

| Models | Coefficient of determination R^2 on Test Data |
|---|---|
| Linear Regression | 0.82798 |
| Decision Tree | 0.834401 |
| Random Forest | 0.863692 |
| Support Vector Machine | 0.768598 |
| XGBoost | 0.849097 |

**Table 4 (a):** Estimation of accuracy for $NH_3$ in different models

| Models | MAE | MSE | RMSE |
|---|---|---|---|
| Linear Regression | 0.0220825 | 0.00296068 | 0.0544121 |
| Decision Tree | 0.0209051 | 0.00292416 | 0.0540755 |
| Random Forest | 0.0164558 | 0.00236551 | 0.0486365 |
| Support Vector Machine | 0.0461031 | 0.00398271 | 0.0631087 |
| XGBoost | 0.019611 | 0.00259721 | 0.0509629 |

**Table 4 (b):** Estimation of error metrics for $NH_3$ in different models

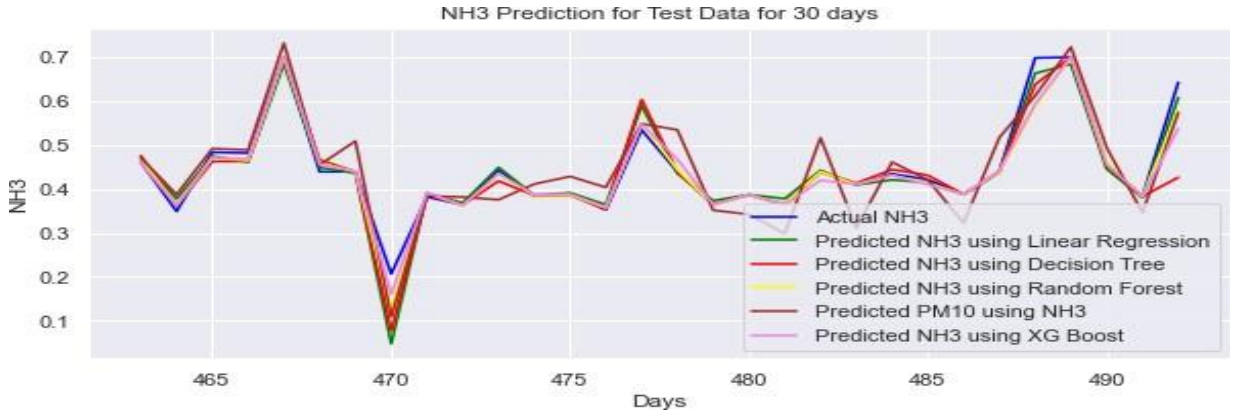


**Fig 4 (a):** Line graph comparing the models' actual and expected values for $NH_3$

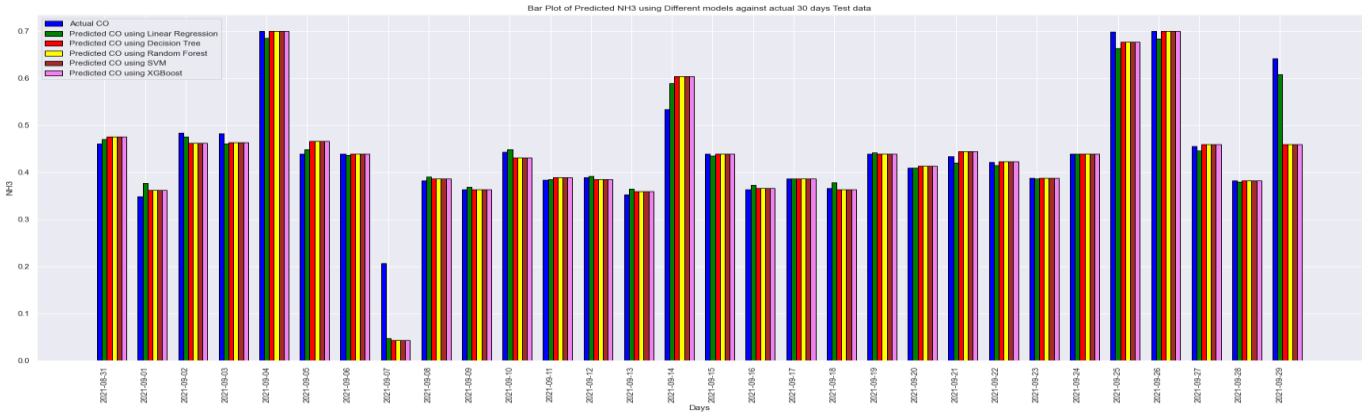


**Fig 4 (b):** Bar graph comparing the models' actual and expected values for $NH_3$

### 4.5. $SO_2$

Estimation of accuracy and error metrics for $SO_2$ in different models

| Models | Coefficient of determination R^2 on Test Data |
|---|---|
| Linear Regression | 0.83404 |
| Decision Tree | 0.770791 |
| Random Forest | 0.885418 |
| Support Vector Machine | 0.813137 |
| XGBoost | 0.883537 |

**Table 5 (a)**: Estimation of accuracy for $SO_2$ in different models

| Models | MAE | MSE | RMSE |
|---|---|---|---|
| Linear Regression | 0.0400133 | 0.00484021 | 0.0695716 |
| Decision Tree | 0.0431597 | 0.0057056 | 0.0755355 |
| Random Forest | 0.0337864 | 0.0033607 | 0.0579716 |
| Support Vector Machine | 0.0542541 | 0.00544984 | 0.073823 |
| XGBoost | 0.0348083 | 0.00339664 | 0.0582807 |

**Table 5 (b):** Estimation of error metrics for $SO_2$ in different models

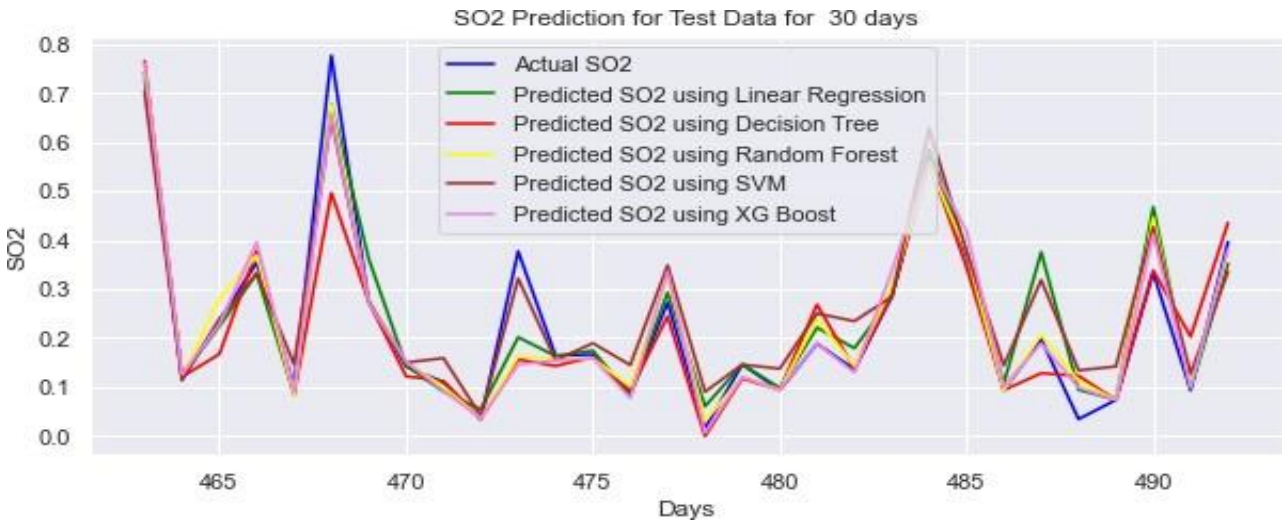


**Fig 5 (a):** A line graph comparing the models' estimated and real $SO_2$ levels

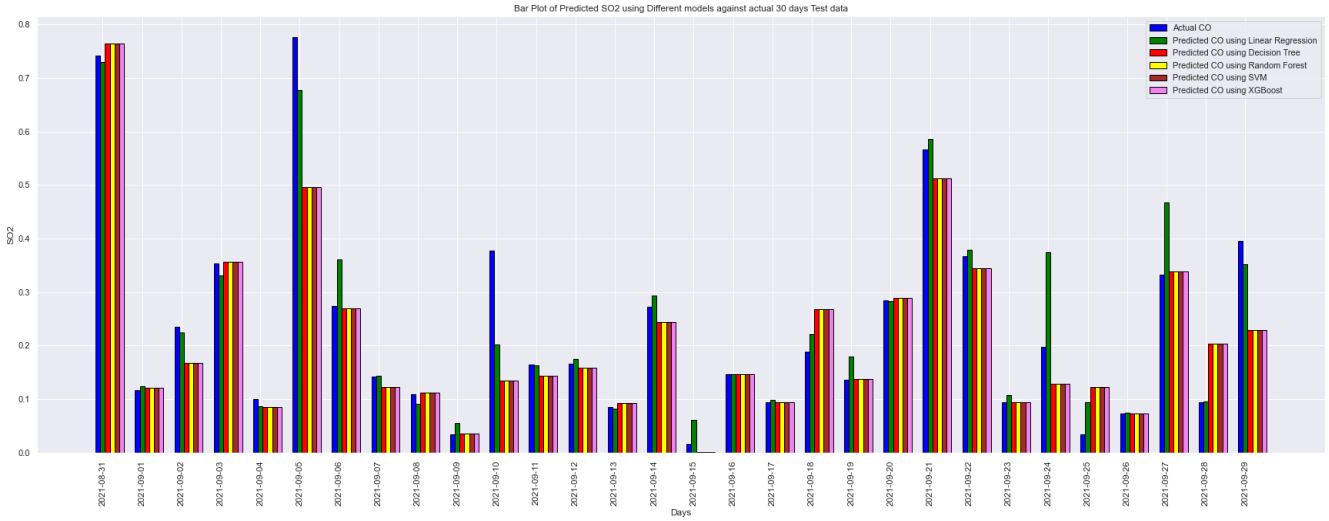


**Fig 5 (b):** Bar graph showing how the models' anticipated and real $SO_2$ levels compare

### 4.6. CO

Assessment of precision and inaccuracy measures for CO in various models

| Models | Coefficient of determination R^2 on Test Data |
|---|---|
| Linear Regression | 0.921752 |
| Decision Tree | 0.848762 |
| Random Forest | 0.928037 |
| Support Vector Machine | 0.872661 |
| XGBoost | 0.907438 |

**Table 6 (a):** Accuracy estimation for CO in various models

| Models | MAE | MSE | RMSE |
|---|---|---|---|
| Linear Regression | 0.0254029 | 0.00170739 | 0.0413206 |
| Decision Tree | 0.0274254 | 0.00309524 | 0.0556349 |
| Random Forest | 0.0187556 | 0.00157835 | 0.0397284 |
| Support Vector Machine | 0.0405885 | 0.00277858 | 0.0527122 |
| XGBoost | 0.0206101 | 0.00201973 | 0.0449414 |

**Table 6 (b):** Estimation of error metrics for CO in different models

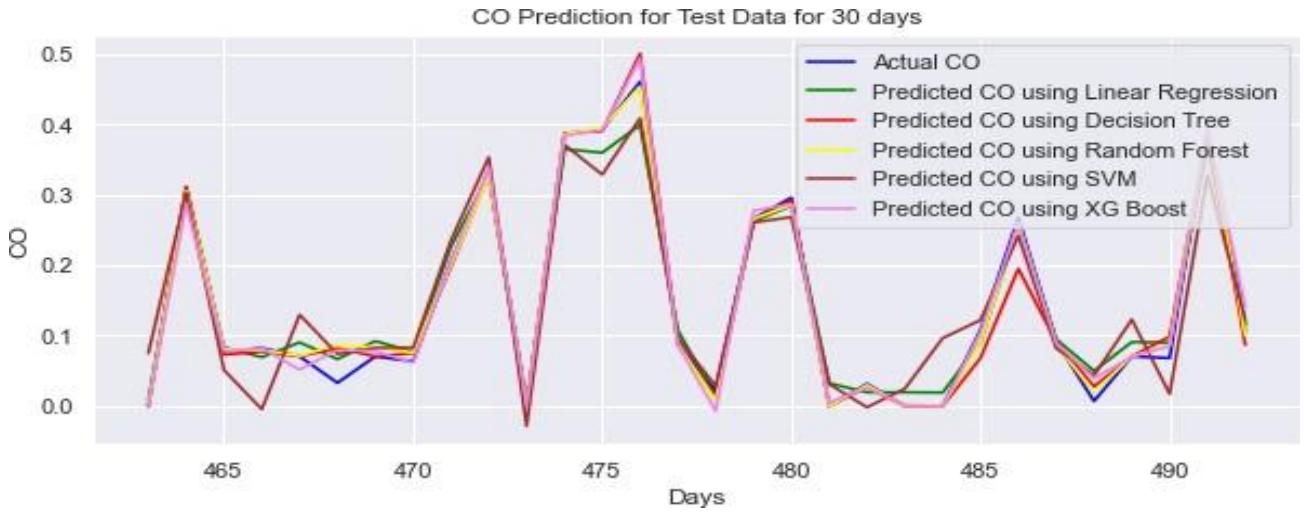


**Fig 6 (a):** A line graph that compares the models' actual and expected CO levels

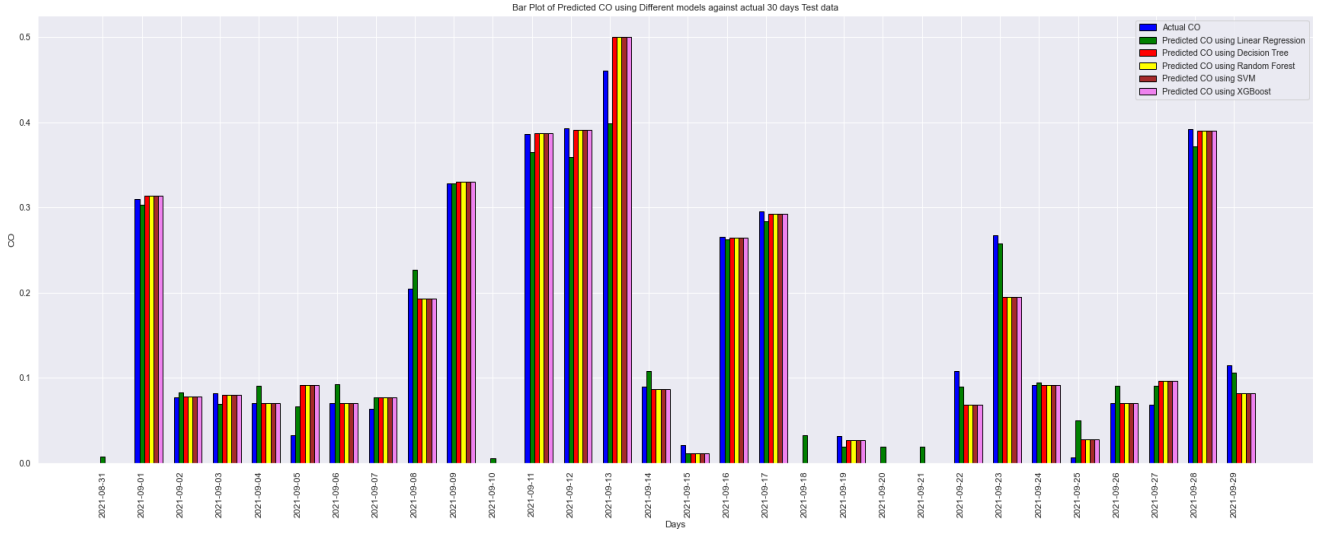


**Fig 6 (b):** Bar graph for comparing the models' actual and expected CO level

### 4.7. $O_3$

Estimation of accuracy and error metrics for $O_3$ in different models

| Models | Coefficient of determination R^2 on Test Data |
|---|---|
| Linear Regression | 0.471247 |
| Decision Tree | 0.151841 |
| Random Forest | 0.488067 |
| Support Vector Machine | 0.41508 |
| XGBoost | 0.458913 |

**Table 7 (a):** Estimation of accuracy for $O_3$ in different models

| Models | MAE | MSE | RMSE |
|---|---|---|---|
| Linear Regression | 0.0592842 | 0.00998107 | 0.0999053 |
| Decision Tree | 0.0717411 | 0.0162444 | 0.127454 |
| Random Forest | 0.0543182 | 0.00952098 | 0.0975755 |
| Support Vector Machine | 0.0705496 | 0.0110413 | 0.105078 |
| XGBoost | 0.0578347 | 0.0102139 | 0.101064 |

**Table 7 (b):** Estimation of error metrics for $O_3$ in different models

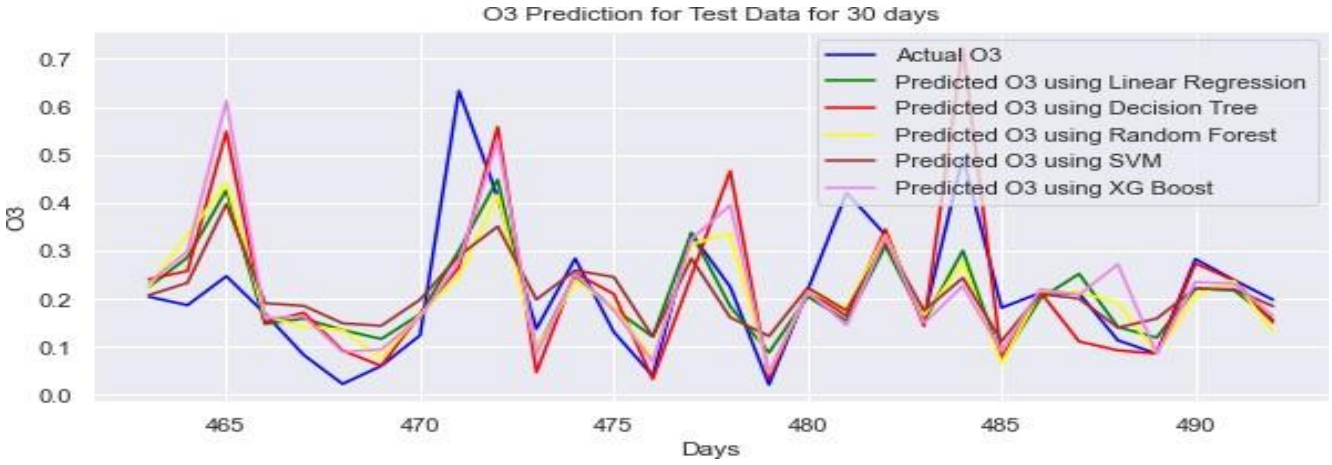


**Fig 7 (a):** A line graph that compares the models' actual and anticipated levels of $O_3$

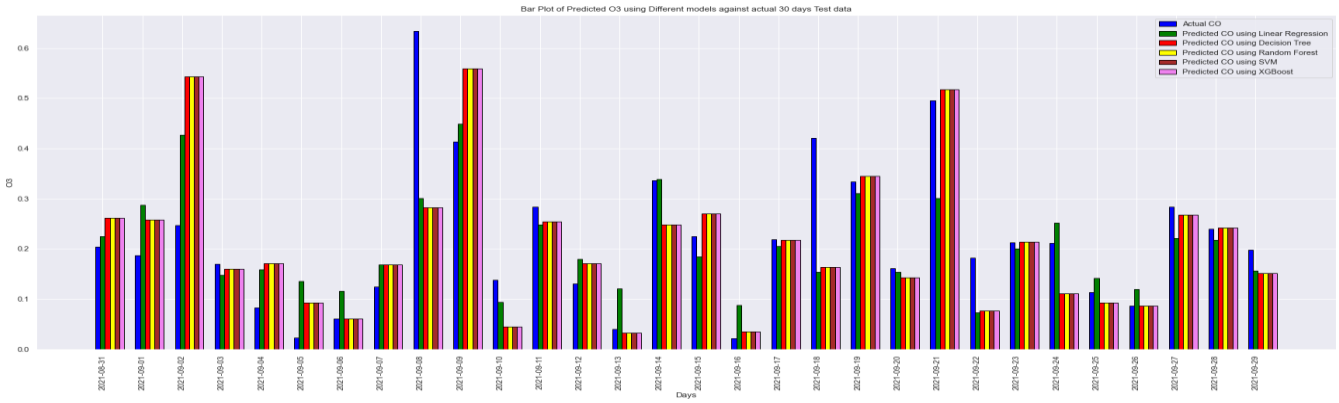


**Fig 7 (b):** Bar graph comparing the models' actual and anticipated levels of $O_3$

Examining the data and graphs for the various models stated above, it is evident that Random Forest has the highest accuracy for all pollutants, with the exception of $NO_2$, for which XGBoost is shown to have the highest accuracy. The results indicate that Random Forest produces the least amount of error for each of the seven contaminants in terms of Mean Absolute Error, or MAE.Once again, in terms of MSE and RMSE, the Random Forest model performs better for six pollutants (NO2 excluded). Regarding the results, XGBoost outperforms the Random Forest model in terms of NO2.

## 5. Conclusions

We have assessed Dhanbad's air quality for the period of April 2019 to March 2023, utilising five key machine learning-based prediction models to analyse pollutants such as $PM_{10}$, NO, $NO_2$, $NH_3$ $SO_2$, CO, and $O_3$. After carefully analysing Dhanbad's air quality rating, our purpose was to compare and contrast five different machine learning methods. Our research leads us to the conclusion that the Random Forest model is most suited for assessing the air pollutants in Dhanbad, namely $PM_{10}$, NO, $SO_2$, $NH_3$, CO, and $O_3$. On the other hand, XGBoost yields the most accurate findings for $NO_2$. The most inaccurate model is the Random Forest.Thus, Dhanbad's air quality is best assessed using the Random Forest model..

## References


[1] O. Bouakline *et al.*, "Prediction of daily PM 10 concentration using machine learning," in *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, IEEE, 2020, pp. 1–5.

[2] K. Tripathi and P. Pathak, "Deep learning techniques for air pollution," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, IEEE, 2021, pp. 1013–1020.

[3] A. Al Yammahi and Z. Aung, "Forecasting the concentration of NO2 using statistical and machine learning methods: A case study in the UAE," *Heliyon*, vol. 9, no. 2, 2023.

[4] S. Peng, J. Zhu, Z. Liu, B. Hu, M. Wang, and S. Pu, "Prediction of Ammonia Concentration in a Pig House Based on Machine Learning Models and Environmental Parameters," *Animals*, vol. 13, no. 1, p. 165, 2022.

*[5]* P. Bhalgat, S. Bhoite, and S. Pitare, "Air quality prediction using machine learning algorithms," *Int. J. Comput. Appl. Technol. Res.*, vol. 8, no. 9, pp. 367–370, 2019.

[6] P. Kadam and S. Vijayumar, "Prediction model: CO 2 emission using machine learning," in *2018 3rd International Conference for Convergence in Technology (I2CT)*, IEEE, 2018, pp. 1–3.

[7] O. A. Ghoneim and B. R. Manjunatha, "Forecasting of ozone concentration in smart city using deep learning," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2017, pp. 1320–1326.

[8] U. K. Sinha, K. Bandyopadhyay and S. C. Dutta, 'Air Quality of Dhanbad During Pre-Lockdown, Lockdown And Post-Lockdown Periods' Journal of Emerging Technologies and Innovative Research (ISSN: 2349-5162), Volume 8, Issue 12, 2021.

[9] S. B. Sonu and A. Suyampulingam, "Linear regression based air quality data analysis and prediction using python," in *2021 IEEE Madras Section Conference (MASCON)*, IEEE, 2021, pp. 1–7.

[10] [10] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[11] U. K. Sinha, K. Bandyopadhayay, and S. C. Dutta, "Prediction of PM10 AND CO Using Support Vector Regression To Analyses The Air Quality of Dhanbad," *Emerg. TRENDS MULTI Discip. Res. Innov.*, p. 1, 2022.

[12] S. Singh, "Prediction of Air Pollution Using Random Forest," *Ann. Romanian Soc. Cell Biol.*, pp. 19314–19322, 2021.

[13] B. Pan, "Application of XGBoost algorithm in hourly PM2. 5 concentration prediction," in *IOP conference series: earth and environmental science*, IOP publishing, 2018, p. 012127.

[14] M. Hussain, S. Afrin, A. Irin, and S. K. Park, "Applying Decision Tree Algorithm for Air Quality Prediction in Bangladesh," in *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*, IEEE, 2021, pp. 1–6.

[15] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006.

[16] D. Parbat and M. Chakraborty, "A python based support vector regression model for prediction of COVID19 cases in India," *Chaos Solitons Fractals*, vol. 138, p. 109942, 2020.

[17] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 03, pp. 90–95, 2007.

[18] M. Waskom, "seaborn: statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[19] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Mach. Learn. PYTHON*.

[20] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, 2021.