Predicting Crop Yield : A Comparative Study of Regression Models

Amiya Ranjan Panda
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
amiya.pandafcs@kiit.ac.in

Manoj Kumar Mishra
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
manojfcs@kiit.ac.in

Lavanya Upadhyay
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
2128077@kiit.ac.in

Sudeshna Rath
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
2128101@kiit.ac.in

Pratyush Kumar Prasad
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
2128035@kiit.ac.in

Kalyanbrata Giri
School of Computer Engineering
KIIT *deemed to be university*
Bhubaneswar, India
2128075@kiit.ac.in

***Abstract*—The purpose of this study is to evaluate how well different machine learning models perform on Kaggle based FAO crop yield dataset which analyzes the crop yields of ten most consumed crops all over the world. Linear Regression, Gradient Boost Regression,Ridge Regression,Lasso Regression, Decision Tree Regression, Random Forest Regression, SVC Regression, K Neighbours Regression, regressors were used to predict the yield of the crop in the dataset. Testing and Training was used to ensure the robustness of the models, and RMSE,MAE and R2 Score along with few other metrics was used as the evaluation metric. The results show that the Random Forest Regressor had the lowest RMSE of 9976.13(approximately), followed by K Neighbours Regression and Decision Tree Regression. The highest RMSE was obtained by SVC with error 93125.93(approximately). Our analysis suggests that the Random Forest Classifier outperforms other algorithms in predicting crop yield within this dataset. This finding holds promise for developing more precise crop yield prediction models in agriculture.**

***Keywords*— Decision Tree Regression,Food and Agriculture Organisation,Gradient Boosting Regression,K Neighbours Regression,Lasso Regression.**

I. INTRODUCTION

Agriculture,the backbone of civilization, faces [1] unprecedented challenges. A growing population demands more food, while environmental alterations disrupt traditional methods. Machine Learning is poised to revolutionize the way we cultivate the land and increase yield and efficiency.

ML algorithms can analyze vast datasets on weather [2] patterns, which allows for precision agriculture. The data-driven approach of incorporating Agriculture and Machine Learning optimizes resource allocation, leading to increased yields and reduced waste, altogether making economic profits for the farmers.

However, challenges stay. Data quality and accessibility are crucial for effective ML models. Additionally, the [3] cost of implementing such technologies can be a barrier for small-scale farmers.

In conclusion, ML offers a transformative vision for agriculture. By harnessing the power of data, it can enhance yields. As we cultivate the future, integrating ML is essential to ensure food security for generations to come.

II. RELATED WORKS

In this, they talk about the analysis, in which they found [4] out the most used features are temperature, rainfall, and soil type, and the most applied algorithm is Artificial Neural Networks in these models.

In this,they discuss a system that uses machine learning [5] techniques to improve farming practices. Their method will recommend the most appropriate crop for a certain piece of land based on weather and content characteristics. Additionally, the system offers details on the necessary amount and composition of fertilizers as well as the seeds needed for cultivation.

Crop production is difficult to predict since it is affected by several factors such as water, ultraviolet (UV), [6] pesticides, fertilizer, and the extent of land covered in that location.

India's enormous population and uncertain climate make it crucial to safeguard global food supplies. When there is a drought, farmers have major challenges. The soil type has both high and low production potential. [7]

Crop yield and crop water productivity (CWP) [8] measurements must be accurate and high-resolution to understand and anticipate spatiotemporal variation in agricultural output capacity.

III. BASIC CONCEPTS

a. *Linear Regression*

It is a statistical approach that models the connection [9] between a dependent variable and one or more independent variables.The model makes the assumption that the independent variables (predictors) and the dependent variable (response) have a linear relationship.Linear regression seeks to identify the best-fitting line (or hyperplane in the case of many predictors) that minimizes the sum of squared discrepancies between observed and predicted values.

In crop yield prediction, linear regression may be used to forecast yields depending on temperature, rainfall, soil nutrients, and agricultural methods.

b. *Gradient Boosting Regression*

Gradient Boosting for regression tasks, is an [10] ensemble machine learning approach. Gradient descent is used to minimize the errors of the preceding model, therefore building an ensemble of weak regression models, usually decision trees, in a sequential fashion. Because of its strong predictive performance and durability, it is frequently employed in a variety of fields to increase accuracy by combining predictions from numerous models.

c. *Ridge Regression*

Ridge regression is a regularization technique used to address multicollinearity in regression models. It [11] includes a penalty term in the ordinary least squares (OLS) objective function to punish big coefficients.The penalty term prevents overfitting by lowering the coefficients to zero, essentially diminishing their size.Ridge regression is very beneficial in agricultural production prediction when dealing with strongly linked predictor variables like temperature and humidity.

d. *Lasso Regression*

Lasso regression, like ridge regression, is a regularization approach that reduces overfitting and improves the interpretability of regression models.It incorporates a penalty component into the OLS [12] objective function, but unlike ridge regression, it employs the L1 norm for regularization.Lasso regression increases sparsity in the coefficient vector, causing some coefficients to be absolutely zero.

This attribute makes lasso regression beneficial for feature selection, since it can automatically discover and pick the most relevant predictors for crop yield prediction.

e. *K Nearest Regression (KNN)*

K Nearest Regression is a supervised machine [13] learning and proximity-based algorithm that uses the proximity of previous data points to predict values or labels for incoming data points. For classification, it assigns the majority class label; for regression, it predicts the average value among the K nearest neighbors. The number of neighbors taken into consideration is determined by the K parameter, which also affects how well KNN performs. Though straightforward and easy to understand, KNN can be computationally taxing, particularly when dealing with big datasets, and may not function well with irrelevant or noisy characteristics. However, it is a good starting model, particularly when used with ensemble methods or on smaller datasets.

f. *Decision Tree Regression*

Decision tree regression is a non-parametric [14] supervised learning technique used for regression tasks.It recursively partitions the feature space into regions where each partition represents a decision based on a function value.Decision tree regression predicts the average target value of an object. as an object. training instances for each leaf node.Decision trees are intuitive and easy to interpret, so they are suitable for [15] performance prediction tasks where explainability is important.

g. *Random Forest Regression*

For classification and regression problems, the ensemble learning technique Random Forest is employed. Using feature randomness and bootstrapped samples of the data, it constructs numerous decision trees. For regression, the final prediction is derived by average over all trees, and for classification, by voting. Random Forest works well with high-dimensional data since it is resilient and less prone to overfitting.

h. *Support Vector Regression*

For regression problems, the supervised machine learning method Support Vector Regression (SVR) [16] is used. In order to minimize error, it locates a hyperplane that optimizes the margin between data points within a certain margin (epsilon). SVR is resistant against outliers and uses kernel functions to address non-linear connections. It is frequently used in finance, economics, and engineering for tasks like function approximation and stock price prediction because it is memory efficient.

IV. Methodology

After importing the required libraries like numpy, seaborn, pandas, used in the prediction model. We initiate by reading the crop yield dataset using the pandas.read_csv function. Once the dataset is loaded, we start exploring and analyzing the dataset.

1. *Exploratory Data Analysis*

In fig 1, we plot the histogram of all the numeric attributes, which give us insight about the data distribution.

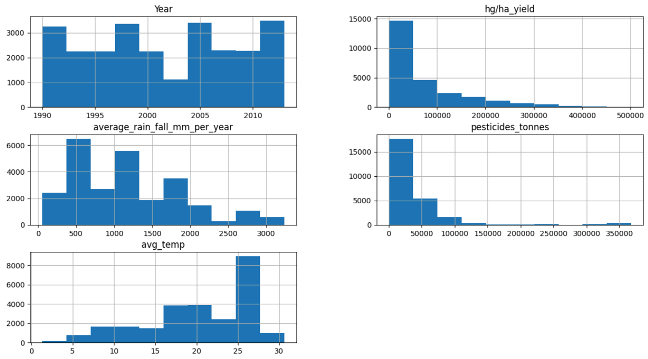


Fig. 1: Attribute histogram

In Fig 2, we plot multiple scatter plots amongst various attributes of the dataset.

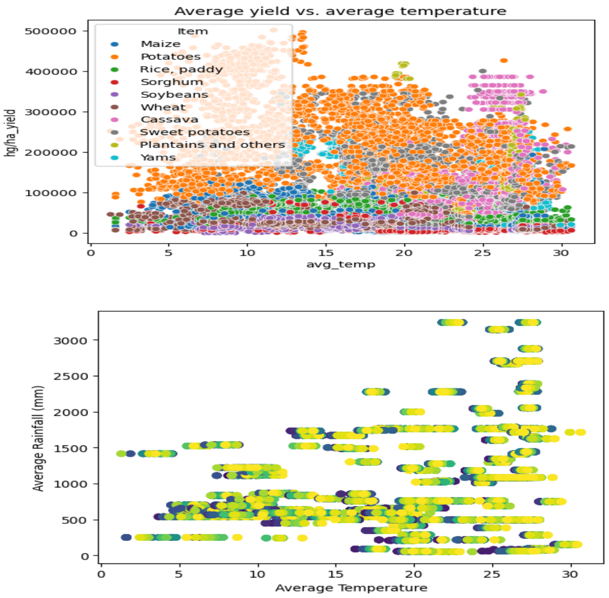


Fig. 2: Scatter plots of Attributes

In Fig 3, A counter plot is plotted to establish the relation between the item count and the area they are produced in as we can see in Fig 3.

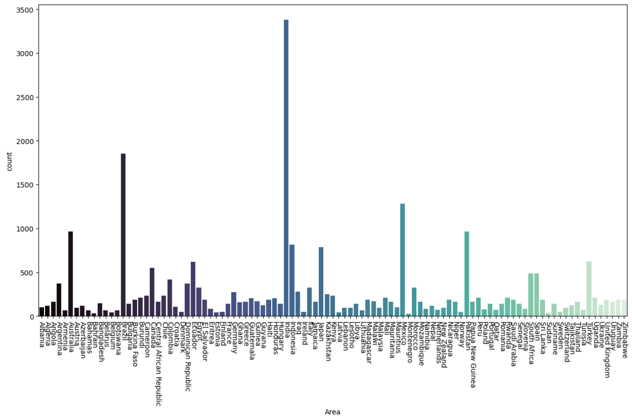


Fig. 3: Counter plot

In Fig 4, Furthermore, we establish the correlation between the items and other attributes, per year, using a **Heatmap**

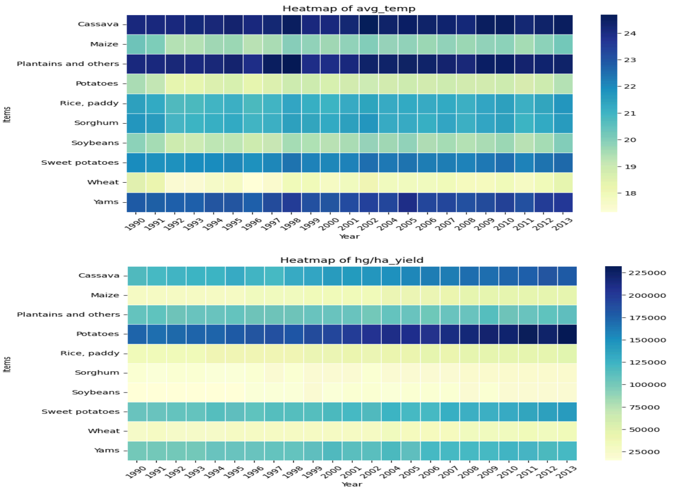
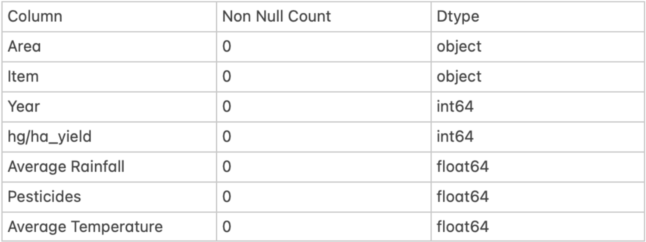


Fig. 4: Heatmap

2. *Data Preprocessing*

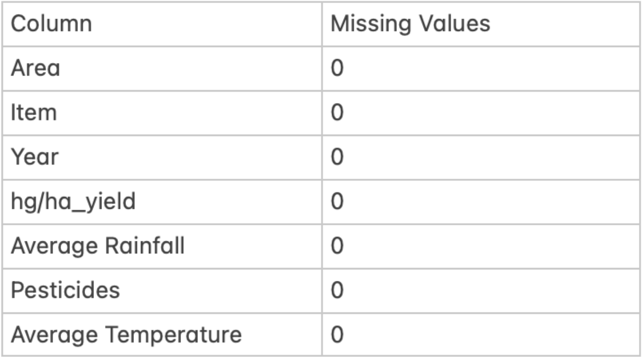
I. Missing Data - in which all the data which are missing are identified and dealt with using various methods like dropping rows with missing / null values, dropping an entire column with missing values.



Column	Non Null Count	Dtype
Area	0	object
Item	0	object
Year	0	int64
hg/ha_yield	0	int64
Average Rainfall	0	float64
Pesticides	0	float64
Average Temperature	0	float64

Fig. 5: missing values

After processing our data for any missing values, we then find the sum of all missing values which turn out to be 0 like shown in fig 5.



Column	Missing Values
Area	0
Item	0
Year	0
hg/ha_yield	0
Average Rainfall	0
Pesticides	0
Average Temperature	0

Fig. 6: sum of missing values

II. Outlier Detection- For the detection of Outliers, we start by plotting Box Plots for each attribute, by which we can visualize the outliers.

We implement the Interquartile Ranges (IQR) for detection of outliers in the dataset used for analysis.

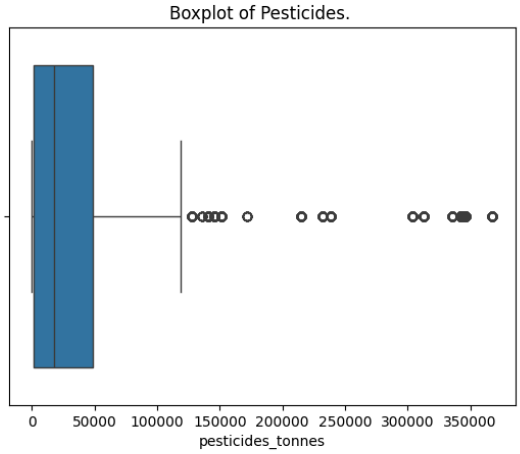


Fig. 7:box plot for detecting outlier

3. *Defining Function*

I. Pearson Correlation- For a regression model, this measures the strength between any two [17] attributes of the dataset

II. Coefficient of Determination (R square)- calculates performance by calculating [18] proportion of variance in dependent variable and independent variable.

III. Mean Squared Error (MSE)- model [19] performance is estimated by averaging the square of errors.

IV. Root Mean Squared Relative Error (RMSRE)- gauging performance of the machine by taking the root of the squared error and then normalizing [20] it by dividing it by the total squared error.

V. Mean Absolute Percentage Error- asiest for interpretation as it is presented by the model's [21] error in percentage.

4. *Models -*

Finishing all of the above stated methods, now we move on to training the Machine using various models. We will apply regression models on our data set.

I. Linear Regression
II. Gradient Boosting Regression
III. Ridge Regression
IV. Lasso Regression
V. k- Nearest Neighbour Regression
VI. Decision Tree Regression
VII. Random Forest Regression
VIII. Support Vector Regression

We train our machine through all the models, one by one, and find out the performance of the model for the dataset, using various evaluation metrics for regression.

V. Result Analysis

Regression metrics are quantitative measures used to evaluate the performance of a regression model. These metrics help you assess how well your model fits the data and predicts future values. Here are the implemented regression metrics.

A. Error Metrics:

- ***Pearson correlation coefficient(R)***: The strength of the linear association between the variables increases with the r score's absolute value becoming closer to 1 (either positively or negatively). The linear link is weaker the closer the r score is to 0.
- ***R Square Score*** : This measure shows how much of the variance in the independent variables (features) accounts for the variance in the dependent variable (target). A higher number on the scale of 0 to 1 denotes a better match.
- **Root Mean Squared Error (RMSE):** This represents the MSE square root. Using the same units as your target variable makes interpretation simpler. A better match is indicated by a lower RMSE.
- **Mean Absolute Error (MAE):** The average absolute difference between the expected and actual values is determined using this statistic. In contrast to MSE, it is less susceptible to outliers. A better match is indicated by a lower MAE.
- ***Mean Absolute Percentage Error (MAPE)***: Better model performance is indicated by lower MAPE values. A MAPE of 0% denotes complete accuracy, meaning that the values predicted by the model match the real ones precisely.IncreasedMAPE values indicate increased prediction errors.

MODELS	**R**	**R2 Score**	**RMSE**	**MAE**	**MAPE**
Linear Regression	0.86451514	0.747312327961836	42630.324526694	29907.491754632363	0.9117460796283101
Ridge Regression	0.86449494	0.7473042337448462	42631.00729908857	29864.88758307408	0.9078287218818285
Lasso Regression	0.86451789	0.7473261756207235	42629.156408211	29893.99762450549	0.9094421813124439

Gradient Boosting Regression	0.97676114	0.9533181810670842	18323.16516975166	10837.581316159703	0.34207906816599815
K -Nearest Neighbours Regression	0.99243046	0.9849021762487353	10420.385530907171	4620.037285521496	0.10057406605524502
Decision Tree Regression	0.98976682	0.9795587859983795	12124.94993528229	3926.90649701176	0.08413478737784882
Random Forest Regression	0.99306713	0.98616207092107	9976.131423430012	3991.2773857721227	0.09420083802563672
Support Vector Regression	0.55635755	-0.20583547930779959	93125.93379849613	57810.26900931254	1.3791109278814748

Table 1 : Error analysis of all implemented models

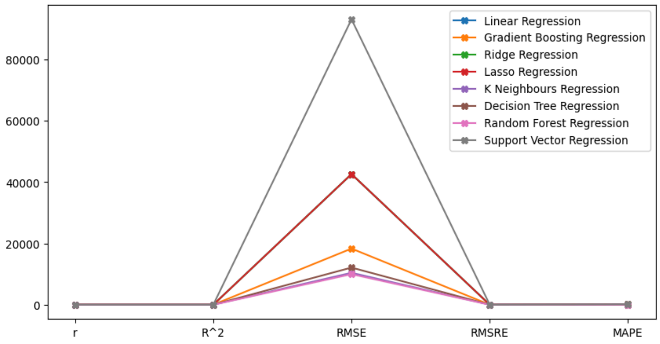


Fig 8: Comparison of error metrics for all models

By looking at Table 1 , it is quite evident that Random Forest Regression performs better than the rest of the models with the least RMSE value and MAE value.

B. ***Random Forest Regressor*** displays a good fit of the model by analyzing the graph of actual versus predicted values as shown in fig 9.

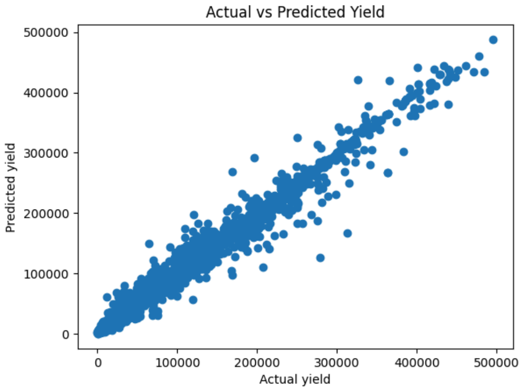


Fig 9 : Actual Yield vs Predicted Yield by Random Forest Regression model

C. ***Residual Analysis*****:** Plotting the residuals (difference between predicted and actual values) can reveal patterns or trends in the errors.
A well-performing model will have residuals scattered randomly around zero (the red dashed line) just like the residual plot for Random Forest Regression as shown in fig 10.

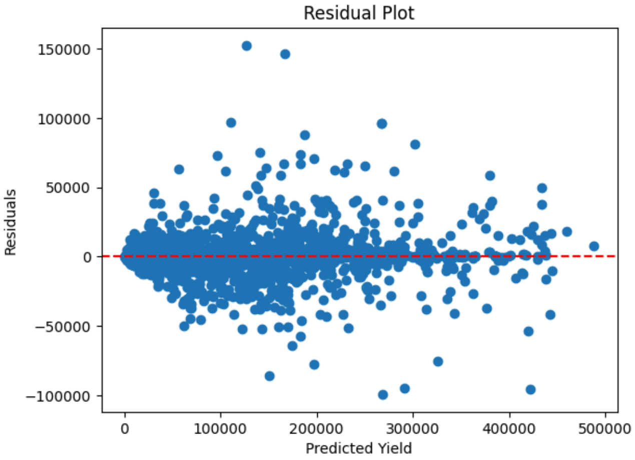


Fig 10 : Residual Plot for Random Forest Regression model

VI. Conclusion & Future Scope

Machine learning algorithms for crop prediction provide a comforting tool for farmers seeking to increase agricultural production. By examining elements like historical data and weather patterns, these models may estimate acceptable crops for a certain geographical area and predict crop yields.This information enables farmers to make more educated decisions regarding planting and resource allocation, eventually enhancing efficiency and profitability.Weather sensors can help improve machine learning models for crops. This not only improves agricultural output projections, but it also reforms farming by placing data in farmers' hands. Prediction models based on machine learning have enormous potential to improve agriculture by encouraging data-driven decision-making, ultimately helping to ensure food supply in the future.

References

[1] Juwono, F. H., Wong, W. K., Verma, S., Shekhawat, N., Lease, B. A., & Apriono, C. (2023, December 1). Machine learning for weed–plant discrimination in agriculture 5.0: An in-depth review. Artificial Intelligence in Agriculture.

[2] Crop Yield Prediction using Machine Learning Algorithm. (2021, May 6). IEEE Conference Publication | IEEE Xplore.

[3] Iniyan, S., Varma, V. A., & Naidu, C. T. (2023, January 1). Crop yield prediction using machine learning techniques. Advances in Engineering Software (1992).

[4] *Crop yield analysis using machine learning algorithms*. (2020, June 1). IEEE Conference Publication | IEEE Xplore.

[5] Varigonda, R. C., Hanmanthgari, S., & Gaddam, R. (2023). CROP YIELD PREDICTION AND EFFICIENT USE OF FERTILIZERS. *International Research Journal of Modernization in Engineering Technology and Science*.

[6] Cheng, M., Jiao, X., Shi, L., Peñuelas, J., Kumar, L., Nie, C., Wu, T., Liu, K., Wu, W., & Jin, X. (2022). High-resolution crop yield and water productivity dataset generated using random forest and remote sensing. *Scientific Data*, *9*(1)

[7] Van Klompenburg, T., Kassahun, A., & Çatal, Ç. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709.

[8] Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021, September 7). *An interaction regression model for crop yield prediction*. Scientific Reports.

[9] Anderson, W. A., Dippel, A. B., Maiden, M. M., Waters, C. M., & Hammond, M. C. (2020). Chemiluminescent sensors for quantitation of the bacterial second messenger cyclic di-GMP. In *Methods in enzymology on CD-ROM/Methods in enzymology* (pp. 83–104).

[10] Otchere, D. A., Ganat, T., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science & Engineering*, *208*, 109244.

[11] Ridge Regression as a Technique for Analyzing Models with Multicollinearity on JSTOR. (n.d.). *www.jstor.org*.

[12] Regression shrinkage and selection via the lasso on JSTOR. (n.d.). *www.jstor.org*.

[13] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbor (KNN) algorithm and its different variants for disease prediction. *Scientific Reports (Nature Publishing Group)*, *12*(1).

[14] Anwla, P. K. (2021, December 1). *Decision Tree Algorithm overview explained*. TowardsMachineLearning.

[15] Nailman, A. (2023, February 12). *Exploring Machine Learning Models: Classification for data analysis*. Machine Learning Models.

[16] Evgeniou, T., & Pontil, M. (2001). Support Vector Machines: Theory and applications. In *Lecture notes in computer science* (pp. 249–257).

[17]Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768.

[18]Filho, D. B. F. Júnior, J. a. S., & Da Rocha, E. C. (2011). What is R2 all about? *Leviathan (São Paulo. Online)*, *3*, 60.

[19]Hodson, T., Over, T. M., & Foks, S. S. (2021). Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems*, *13*(12).

[20]Hodson, T. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487.

[21]Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679.