Journal of Agrometeorology

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 26 (3) : 344-348 (September - 2024)
https://doi.org/10.54386/jam.v26i3.2598
https://journal.agrimetassociation.org/index.php/jam

Research Paper

Weather based paddy yield prediction using machine learning regression algorithms

DHINAKARAN SAKTHIPRIYA[1,] and THANGAVEL CHANDRAKUMAR*[1]

[1]Department of Applied Mathematics & Computational Science, Thiagarajar College of Engineering, Madurai 625 015, Tamil Nadu.
**Corresponding Author: t.chandrakumar@gmail.com*

ABSTRACT

Paddy is a major crop in India which is highly affected by the weather variables resulting in drastic reduction of its yield; adverse all the variables drastically reduce the paddy yield. In this research, machine learning model was developed for prediction of paddy yield production by linear regression (LR), random forest regression (RFR), support vector regression (SVR), cat boost regression (CBR), and hybrid machine learning with variance inflation factor (VIF) LR-VIF, RFR-VIF, SVR-VIF, and CBR-VIF techniques. The dataset consists of variables (weather) for more than 15 years collected for the study area which is Madurai district, Tamil Nadu in India. Analysis was carried out by fixing 70% of data calibration & remaining 30% for validation in Jupyter notebook (Python programming). Results showed that CBR-VIF performed having nRMSE value 1.23 to 1.40% for Madurai South, nRMSE value 0.56 to 1.40% for Melur, nRMSE value 1.10 to 1.25% for Usilampatti, and nRMSE value 0.75 to 1.10% for Thirumangalam. The hybrid model of CBR along with VIF and then CBR model has shown improvement with high influenced weather variables such as maximum temperature, minimum temperature, rainfall normal, and actual rainfall.

Keywords: Paddy seed, Hybrid machine learning model, Linear regression (LR), Random Forest regression (RFR), Support vector regression (SVR), Cat boost regression (CBR).

The paddy crop serves as an essential component of economic survival and as the foundation for international assistance. Farmers encounter numerous challenges caused by variables including water scarcity, price volatility resulting from supply and demand dynamics, unpredictability of weather conditions, soil nutrient deficiencies, and imprecise crop forecasts (Shankar *et al.,* 2022; Saravanan and Bhagavathiappan, 2022; Joshua *et al.,* 2022). The estimation of agricultural yields, particularly paddy, is a complex undertaking due to its reliance on a variety of elements including lineage, environmental conditions, farming techniques, and the interplay between them (Joshua *et al.,* 2021; Zhou and Ismaeel 2021; Sridhara *et al.,* 2024; Leng and Hall, 2020; Elbasi *et al.,* 2023). Various research works have been carried out using machine learning algorithms for crop yield prediction, rice cultivar quality measurement, soil conditions, fertilizers, prediction of rice cultivar for many more crops globally. For example, Ekanayake *et al.,* (2021) development of crop-weather models for the paddy yield in Sri Lanka based on nine weather indices, including rainfall, relative humidity (minimum and maximum), temperature (minimum and maximum), wind speed (morning and evening), evaporation, and sunlight hours. Using random forest (RF) and found that minimum relative humidity and the maximum temperature are the most significant weather indicators for paddy cultivation. Rakhee *et al.,* (2018) proposed the development of fuzzy regression models to forecast rice yield in the Kanpur district. Setiya and Nain (2021) presented a regression model that effectively forecasts rice crop yield in the USN district by utilizing the variability of rainfall, minimum temperature, maximum temperature, and solar radiation. For the crops like millets, groundnut, wheat, sugarcane, rice, cotton and coriander various researches were carried out by the researchers with soil characteristics (texture, pH, color, permeability, drainage, water retention, and erosion), temperature, rainfall, humidity, sunlight using machine learning models such as (Random Tree, K-Nearest Neighbor, polynomial regression, Random Forest, Artificial Neural Network, and Support Vector Regression and Naïve Bayes) for yield predictions in all the regions of India (Kumar *et al.,* 2019; Pudumalar *et al.,* 2017; Nischitha *et al.,* 2020; Ali *et al.,* 2021; Kumar *et al.,* 2014; Krithika *et al.,* 2022).

Present study was undertaken to develop the prediction

Article info - DOI*: *https://doi.org/10.54386/jam.v26i3.2598
Received: 15 May 2024; Accepted: 18 July 2024 ; Published online : 01 September 2024

"This work is licensed under Creative Common Attribution-Non Commercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) © Author (s)"

Table 1: Divisions of Madurai district under study

S.No	Division name	Latitude	Longitude
1	Madurai south	9.9252° N	78.1198° E
2	Melur	10.0304° N	78.3438° E
3	Usilampatti	9.9597° N	77.8007° E
4	Thirumangalam	9.8221° N	77.9828° E

models for paddy yield in Madhurai district of Tamil Nadu using various hybrid machine learning techniques.

MATERIALS AND METHODS

Data collection

The geographical area of the Madurai District, which is between 9.32^0 and 10.18^0 North Latitude and 77.28^0 and 78.27^0 East Longitude, is the subject of this study. It encompasses 3,74,173 hectares (or 3,741,73 square kilometers). The Madurai district is divided into four divisions: Madurai, Melur, Usilampatti, and Thirumangalam. This data includes daily weather information, compositions during the paddy seed growing period for all four divisions (Table 1).

For this research, extensive data over the past fifteen years (2007-2022) have been collected from the Tamil Nadu Agricultural University, soil test laboratory, agriculture engineering department for all divisions in Madurai, Tamil Nadu. The dataset was sourced from the website: http://tawn.tnau.ac.in/General/DistrictWiseSummaryPublicUI.aspx?RW=1. This data includes daily weather information, soil nutrient compositions, and water availability during the paddy seed growing period for all four divisions: Madurai, Melur, Usilampatti, and Thirumangalam. The machine learning algorithms extracted data features, traits, or input variables (predictors), while the intended result or prediction was represented by the output variable, also known as the target variable or label. Table 2 lists the variables used as inputs (predictors) and outputs (targets) in this study. The research involves constructing multiple ML regression algorithms for the dataset, with computed prediction limits serving as targets.

Development of models

The various machine learning techniques viz. Linear Regression (LR), Random Forest Regression (RFR), Support Vector Regression (SVR), Cat Boost Regression (CBR), and hybrid machine learning LR-VIF (Linear Regression with Variance Inflation Factor), RFR-VIF (Random Forest Regression with Variance Inflation Factor) SVR-VIF (Support Vector Regression with Variance Inflation Factor) and CBR-VIF (Cat Boost Regression with Variance Inflation Factor), were applied for Madurai districts, (Madurai south, Melur, Usilampatti, and Thirumangalam). Python with jupyter notebook was used for implementing the machine learning model for developing the paddy yield production with the evaluation metrics such as R^2, MSE, RMSE, and nRMSE. For hybrid machine learning models combination of LR with LR-VIF, RFR with RFR-VIF, SVR with SVR-VIF and CBR with CBR-VIF was done. In the LR-VIF model, the variables are selected by VIF techniques and used as an input variable for LR. In the RFR-VIF model, variables are selected by VIF techniques and these variables are used as an input variable for RFR. In the SVR-VIF model, variables are selected by VIF techniques and these variables are used as an input variable for SVR. In the CBR-VIF model, variables are selected by VIF techniques and these variables are used as an input variable for CBR. Ten years (2007-2017) data of various weather variables were used for training and 5 years (2018-2022) of data were used for testing.

Model accuracy

The effectiveness of statistical models was evaluated by computing the coefficient of determination (R^2), mean squared error (MSE), root mean square error (RMSE), and normalized root mean square error (nRMSE) employing the following formula.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Ai - Fi)^2}{\sum_{i=1}^{N}(Ai - M)^2} \qquad MSE = \frac{1}{N}\sum_{i=1}^{N}(Ai - Fi)^2$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Ai - Fi)^2} \qquad nRMSE = \frac{100}{M} * \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Ai - Fi)^2}$$

where forecast value, actual value, number of observations, and mean of observed value are denoted by Ai, Fi, N, and M, respectively. For the subsequent paddy yield production procedure, nRMSE is categorized as excellent- when its value is below 10%, good- when it lies between 10% and 20%, fair- when it ranges from 20% to 30%, and poor- when it exceeds 30%.

Table 2: Data set used in the study

S. No	Variable name	Variable ID	Variable type	Description
1	Maximum temperature	MAXT	Predictor	Maximum temperature for various division
2	Minimum temperature	MINT	Predictor	Minimum temperature for various division
3	Rainfall normal	RN	Predictor	Rainfall normal value
4	Actual rainfall	AR	Predictor	Rainfall actual value
5	Starting month	SM	Predictor	Starting month for various season
6	Ending month	EM	Predictor	Ending month for various season
7	Division name	DN	Predictor	District list in Madurai district
8	Duration	DUR	Predictor	Duration based on no.of. days
9	Production	PRD	Target	Production ratio
10	Crop year	CY	Predictor	Year of crop production
11	Seed name	SN	Predictor	Collection of paddy name in Madurai districts

Table 3: Performance of paddy yield production by different models for Madurai south

Models	Modal accuracy during calibration (2007 – 2017)				Modal accuracy during validation (2018 – 2022)		
	R^2	MSE	RMSE	nRMSE	MSE	RMSE	nRMSE
LR	0.78	0.054	0.784	4.95	0.055	0.322	10.00
RFR	0.82	0.102	0.156	3.21	0.044	0.218	7.56
SVR	0.74	0.152	0.489	3.62	0.085	0.260	8.15
CBR	0.86	0.015	0.324	1.75	0.019	0.102	3.26
LR-VIF	0.80	0.051	0.568	2.36	0.051	0.328	14.02
RFR-VIF	0.85	0.021	0.478	2.15	0.036	0.180	7.56
SVR-VIF	0.79	0.076	0.325	3.12	0.085	0.156	6.23
CBR-VIF	0.95	0.008	0.052	1.40	0.009	0.076	1.23

Table 4: Performance of paddy yield production by different models for Melur

Models	Modal accuracy during calibration (2007 – 2017)				Modal accuracy during validation (2018 – 2022)		
	R^2	MSE	RMSE	nRMSE	MSE	RMSE	nRMSE
LR	0.63	0.560	0.756	4.99	0.090	0.556	5.12
RFR	0.89	0.145	0.881	5.05	0.042	0.898	3.55
SVR	0.82	0.520	0.620	3.78	0.054	0.260	6.20
CBR	0.94	0.100	0.208	2.56	0.008	0.208	3.66
LR-VIF	0.61	0.350	0.652	3.23	0.087	0.666	2.32
RFR-VIF	0.88	0.870	0.280	3.99	0.069	0.774	5.11
SVR-VIF	0.87	0.580	0.455	4.50	0.063	0.254	2.89
CBR-VIF	0.96	0.050	0.188	1.40	0.007	0.188	0.56

RESULT AND DISCUSSION

Different regression models for paddy yield production by Madurai South

The machine learning models (LR, RFR, SVR, CBR) was tested individually as well as combined with VIF in Hybrid mode (LR-VIF, RFR-VIF, SVR-VIF, CBR-VIF) for the weather variables and the results obtained for the models during calibration and validation for paddy yield prediction for Madurai south are shown in Table 3. All the models proposed in this research exhibited excellent performance with nRMSE value <10. While taking weather variables for analysis, nRMSE value during calibration was between 1.40 to 4.95% and between 1.23 to 14.02% during validation. It is found that the CBR-VIF combination of (CBR and VIF) is having the lowest nRMSE value for the weather variables followed by CBR, SVR-VIF, RFR-VIF, RFR, LR, SVR, and LR-VIF. Based on model performance for paddy yield prediction using weather variables, CBR-VIF followed by CBR were found to be best for Madurai south.

Different regression models for paddy yield production by Melur

The machine learning models (LR, RFR, SVR, CBR) was tested individually as well as combined with VIF in Hybrid mode (LR-VIF, RFR-VIF, SVR-VIF, CBR-VIF) for the weather variables and the results obtained for the models during calibration and validation for paddy yield prediction for Melur are shown in Table 4. All the models proposed in this research exhibited excellent performance with nRMSE value <10. While taking weather variables for analysis, nRMSE value during calibration was between 1.40 to 5.05% and between 0.56 to 6.20% during validation. It is found that the CBR-VIF combination of (CBR and VIF) is having the lowest nRMSE value for the weather variables followed by LR-VIF, SVR-VIF, RFR, CBR, LR, RFR-VIF, and SVR. Based on model performance for paddy yield prediction using weather variables, CBRVIF followed by LRVIF were found to be best for Melur.

Different regression models for paddy yield production by Usilampatti

The machine learning models (LR, RFR, SVR, CBR) was tested individually as well as combined with VIF in Hybrid mode (LR-VIF, RFR-VIF, SVR-VIF, CBR-VIF) for the weather variables and the results obtained for the models during calibration and validation for paddy yield prediction for Usilampatti are shown in Table 5. All the models proposed in this research exhibited excellent performance with nRMSE value <10. While taking weather variables for analysis, nRMSE value during calibration was between 1.25 to 5.84% and between 1.10 to 7.50% during validation. It is found that the CBR-VIF combination of (CBR and VIF) is having the lowest nRMSE value for the weather variables followed by CBR, SVR-VIF, LR-VIF, RFR-VIF, RFR, LR and SVR. Based on model performance for paddy yield prediction using weather variables, CBR-VIF followed by CBR were found to be best for Usilampatti.

Table 5: Performance of paddy yield production by different models for Usilampatti

Models	Modal accuracy during calibration (2007 – 2017)				Modal accuracy during validation (2018 – 2022)		
	R^2	MSE	RMSE	nRMSE	MSE	RMSE	nRMSE
LR	0.60	0.147	0.364	4.58	0.123	0.541	5.01
RFR	0.72	0.358	0.888	5.84	0.211	0.789	5.25
SVR	0.80	0.025	0.478	1.63	0.099	0.361	7.50
CBR	0.90	0.021	0.108	3.75	0.066	0.444	1.56
LR-VIF	0.63	0.247	0.211	3.23	0.111	0.451	3.23
RFR-VIF	0.76	0.369	0.987	2.58	0.255	0.114	4.21
SVR-VIF	0.81	0.045	0.451	4.21	0.095	0.477	2.56
CBR-VIF	0.93	0.018	0.045	1.25	0.015	0.120	1.10

Table 6: Performance of paddy yield production by different models for Thirumangalam

Models	Modal accuracy during calibration (2007 – 2017)				Modal accuracy during validation (2018 – 2022)		
	R^2	MSE	RMSE	nRMSE	MSE	RMSE	nRMSE
LR	0.55	0.111	0.257	6.95	0.102	0.478	6.00
RFR	0.62	0.224	0.200	6.23	0.258	0.389	3.21
SVR	0.88	0.197	0.478	4.44	0.780	0.457	5.22
CBR	0.93	0.100	0.346	2.65	0.121	0.120	2.11
LR-VIF	0.59	0.147	0.158	1.56	0.125	0.124	5.66
RFR-VIF	0.72	0.325	0.302	4.51	0.690	0.210	2.23
SVR-VIF	0.89	0.158	0.178	2.89	0.870	0.255	2.32
CBR-VIF	0.97	0.065	0.150	1.10	0.052	0.100	0.75

Different regression models for paddy yield production by Thirumangalam

The machine learning models (LR, RFR, SVR, CBR) was tested individually as well as combined with VIF in Hybrid mode (LR-VIF, RFR-VIF, SVR-VIF, CBR-VIF) for the weather variables and the results obtained for the models during calibration and validation for paddy yield prediction for Thirumangalam are shown in Table 6. All the models proposed in this research exhibited excellent performance with nRMSE value <10. While taking weather variables for analysis, nRMSE value during calibration was between 1.10 to 6.95% and between 0.75 to 6.00% during validation. It is found that the CBRVIF combination of (CBR and VIF) is having the lowest nRMSE value for the weather variables followed by CBR, RFR-VIF, SVR-VIF, LR-VIF, RFR, SVR and LR. Based on model performance for paddy yield prediction using weather variables, CBR-VIF followed by CBR were found to be best for Thirumangalam.

CONCLUSION

Eight machine learning models were developed to yield the paddy production using weather variables data. The findings indicate that CBR-VIF exhibited the highest performance followed by CBR. The effectiveness of the CBR model was enhanced through hybrid machine learning techniques. In comparison with LR, LRVIF, RFR, RFRVIF, SVR, SVRVIF, CBR, CBRVIF, and SMLR-SVR, CBR-VIF demonstrated superior accuracy in multivariate paddy seed selection. It is also found that the following weather variable (maximum temperature, minimum temperature, rainfall normal, actual rainfall) has influenced the paddy yield. Consequently, considering its superior performance across the area of study, CBR-VIF stands out as a promising choice for district-level multivariate paddy yield production across various locations within Madurai district, Tamil Nadu.

ACKNOWLEDGEMENT

The authors express their gratitude to the Thiagarajar College of Engineering (TCE) for supporting us to carry out this research work. Also, the financial support from TCE under Thiagarajar Research Fellowship scheme (File.no: TRF/Jan-2022/04) is gratefully acknowledged.

Conflict of interests: The authors declare that there is no conflict of interest related to this article.

***Data availability statement*:** To be provided on request.

Authors contribution: **Sakthipriya**: Data collection, Data Analysis, Conceptualization, Methodology, Visualization; **Chandrakumar**: Resources, Supervision, Methodology, Writing-original draft, Writing- review and editing.

Disclaimer: The contents, opinions and views expressed in the research article published in the Journal of Agrometeorology are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

Publisher's Note: The periodical remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

Ali, S. M., Das, B. and Kumar, D. (2021). Machine learning based crop recommendation system for local farmers of

Pakistan. *Revista Geintec-Gestao Inovacao E Tecnologias*, 11(4): 5735-5746.

Ekanayake, P., Rankothge, W., Weliwatta, R. and Jayasinghe, J. W. (2021). Machine learning modelling of the relationship between weather and paddy yield in Sri Lanka. *J. Maths.,* 2021(1): 9941899.

Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E. and Saker, L. (2023). Crop prediction model using machine learning algorithms. *Applied Sci.,* 13(16): 9288.

Joshua, V., Priyadharson, S. M., and Kannadasan, R. (2021). Exploration of machine learning approaches for paddy yield prediction in eastern part of Tamilnadu. *Agronomy*, 11(10): 2068.

Joshua, S. V., Priyadharson, A. S. M., Kannadasan, R., Khan, A. A., Lawanont, W., Khan, F. A., ... and Ali, M. J. (2022). Crop yield prediction using machine learning approaches on a wide specrum. *Comp., Materials Continua*, 72(3): 5663-5679.

Krithika, K. M., Maheswari, N. and Sivagami, M. (2022). Models for feature selection and efficient crop yield prediction in the groundnut production. *Res. Agric. Eng.,* 68(3), 131-141.

Kumar, N., Pisal, R. R., Shukla, S. P. and Pandey, K. K. (2014). Crop yield forecasting of paddy, sugarcane and wheat through linear regression technique for south Gujarat. *Mausam*, 65(3): 361-364.

Kumar, S., Attri, S. D. and Singh, K. K. (2019). Comparison of Lasso and stepwise regression technique for wheat yield prediction. *J. Agrometeorol.,* 21(2): 188-192. https://doi.org/10.54386/jam.v21i2.231

Leng, G. and Hall, J. W. (2020). Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. *Environmental research letters: ERL* [Web site], 15(4): 044027.

Nischitha, K., Vishwakarma, D., Ashwini, M. N. and Manjuraju, M. R. (2020). Crop prediction using machine learning approaches. *Intern. J. Engg. Res. Techn. (IJERT)*, 9(08):23-26.

Pudumalar, S., Ramanujam, E., Rajashree, R. H., Kavya, C., Kiruthika, T. and Nisha, J. (2017). Crop recommendation system for precision agriculture. In 2016 eighth international conference on advanced computing (ICoAC) (pp. 32-36). IEEE.

Saravanan, K. S. and Bhagavathiappan, V. (2022). A comprehensive approach on predicting the crop yield using hybrid machine learning algorithms. *J. Agrometeorol.,* 24(2):179-185. https://doi.org/10.54386/jam.v24i2.1561

Setiya, P. and Nain, A.S. (2021). Development of yield prediction model of rice crop for hilly and plain terrains of Uttarakhand. *J. Agrometeorol.*, 23(4):452-456. https://doi.org/10.54386/jam.v23i4.162

Shankar, T., Malik, G. C., Banerjee, M., Dutta, S., Praharaj, S., Lalichetti, S., ... and Hossain, A. (2022). Prediction of the effect of nutrients on plant parameters of rice by artificial neural network. *Agron.,* 12(9): 2123.

Rakhee, Singh, A. and Kumar, A. (2018). Weather based fuzzy regression models for prediction of rice yield. *J. Agrometeorol.,* 20(4): 297-301. https://doi.org/10.54386/jam.v20i4.569

Sridhara, S., Soumya, B. and Kashyap, G. R. (2024). Multistage sugarcane yield prediction using machine learning algorithms. *J. Agrometeorol.,* 26(1): 37-44. https://doi.org/10.54386/jam.v26i1.2411

Zhou, Q. and Ismaeel, A. (2021). Integration of maximum crop response with machine learning regression model to timely estimate crop yield. *Geo-spatial Information Sci.,* 24(3): 474-483.