Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 60(4): 35-44 (2023)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
https://doi.org/10.53560/PPASA(60-4)820

Pakistan Academy of Sciences

Research Article

Measuring the Performance of Supervised Machine Learning Algorithms for Optimizing Wheat Productivity Prediction Models: A Comparative Study

Malik Muhammad Hussain[1], Farrukh Shehzad[2], Muhammad Islam*[3], Ashique Ali Chohan[4], Rashid Ahmed[2], and H.M. Muddasar Jamil Shera[3]

[1]Department of Statistics, Emerson University Multan, Pakistan
[2]Department of Statistics, the Islamia University of Bahawalpur, Punjab, Pakistan
[3]Crop Reporting Service, Agriculture Department, Punjab, Pakistan
[4]Department of Energy and Environment, Faculty of Agricultural Engineering, Sindh Agriculture University Tando Jam, Pakistan

Abstract: The issue of precise crop prediction gained worldwide attention in the midst of food security concerns. In this study, the efficacies of different machine learning (ML) algorithms, i.e., multiple linear regression (MLR), decision tree regression (DTR), random forest regression (RFR), and support vector regression (SVR) are integrated to predict wheat productivity. The performances of ML algorithms are then measured to get the optimized model. The updated dataset is collected from the Crop Reporting Service for various agronomical constraints. Randomized data partitions, hyper-parametric tuning, complexity analysis, cross-validation measures, learning curves, evaluation metrics and prediction errors are used to get the optimized model. ML model is applied using 75% training dataset and 25% testing datasets. RFR achieved the highest R^2 value of 0.90 for the training model, followed by DTR, MLR, and SVR. In the testing model, RFR also achieved an R^2 value of 0.74, followed by MLR, DTR, and SVR. The lowest prediction error (P.E) is found for the RFR, followed by DTR, MLR, and SVR. K-Fold cross-validation measures also depict that RFR is an optimized model when compared with DTR, MLR and SVR.

Keywords: Model Optimizations, Machine Learning Algorithms, Prediction Models, Performance Measurement.

1. INTRODUCTION

Data science is an interdisciplinary field that comprises scientific computing, processes, algorithms and systems to extract meaningful insights and knowledge from data [1, 2]. Machine learning (ML) is viewed as an advanced tool of data science and artificial intelligence and focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit programming [3, 4]. Arthur Samuel, a pioneer in ML, defined ML as "field of study that gives computers the ability to learn without being explicitly programmed", and used to learn the machines, how to handle the data in more efficient way [5, 6]. ML algorithms created new dimensions and possibilities for data intensive research in the applied agricultural science [7-9]. Various supervise machine learning models, i.e., linear regression, decision tree, random forest and artificial neural networks, etc., are used for the prediction of parameters using agricultural constrains [10, 11].

Global food supply system is facing challenges due to rapid population growth and unpredictable changes in climate [12, 13]. Agriculture being core component, got key significance in the midst of food security situation in the world, and the focus of the agricultural research is diverted towards the improvement and true prediction of crop productivity [14-16]. The high growth rate of population, increases the demand of staple food crop [17]. Wheat, maize and rice are the world's significant cereals crops in terms of area and food supply channels [18]. Wheat is important food crop

Received: May 2023; Accepted: December 2023
* Corresponding Author: Muhammad Islam <mislam6667@gmail.com>

in the developing world after rice [19, 20]. Wheat crop, ranks first in acreage and production among all others food crops [21].

Chlingaryan *et al.*, [22] conducted a study using various machine learning techniques to predict the crop yield, and they concluded that ML approaches provide comprehensive and effective solutions for better crop projection. Bondre and Mahagaonkar [23] studied the efficacies of support vector and random forest machine learning algorithms to predict the crop yield using agricultural dataset of fertilizer application. They demonstrated that random forest is better choice to predict the crop yield.

Cedric *et al.*, [24] proposed prediction system using machine learning algorithms to predict the crops yield in African countries using weather, climatic and chemical datasets. They applied decision tree, multivariate regression, k-nearest neighbor models. They performed hyper-parameter tuning and cross-validation to optimize the model performance to avoid over fitted model. The results indicated that decision tree model is better fitted model. Mohapatra and Chaudhary [25] applied machine learning models, i.e., multiple linear regression, decision tree regression, random forest regression, k-nearest neighbor approach and support vector regression. They applied evaluation metrics to measure the model performances and reported that random forest regression performed well.

Currently, Food supply system in the world is facing major stress due to food insecurity. The advancement in agricultural science and technology leaded immense volume of data. Machine learning algorithms have been emerged as advanced tools of artificial intelligence and applied as an alternative of traditional statistical modeling, to extract meaningful information inside from data using algorithms. The optimization and validations of machine learning algorithms is one of the key concerns for the world, to predict the true agricultural productions. This research will focus, to apply and to integrate different machine learning algorithms such as multiple regression, decision tree, random forest and support vector regression with the aim to search out, which one is best, and able to validate and to optimize the learning algorithms, to predict the wheat productivity in Pakistan using various agronomical constrains.

2. MATERIALS AND METHODS

2.1 Data Collection and Variables Features

The secondary cross-sectional data of 15000 crop cut experiments (CCE), conducted by the Crop Reporting Service is collected along with different agronomical constrains comprise from year 2019 to 2021. This data is further pre-processed and centralized using the centroid clustering scheme introduced by Islam and Shehzad [14] for the 136 wheat area sown tehsils zones in Punjab. The experiment is performed using the python' key library called Scikit Learn (https://scikit-learn.org/stable/supervised_learning.html). The wheat average yield in mds/acre is used as depended (labeled) variable along with agronomical constrains, i.e., urea, DAP and other fertilizers used in kg/acre, water, spray pest, soil type (chikny loom or not), adoption of advanced varieties trend (yes/no), wheat harvested from April 1 to 20 (yes/no), and wheat sown in November (yes/no).

2.2 Hierarchy of Machine Learning Algorithms

The layout hierarchy of supervised machine learning model deployment fallows under the following seven prominent steps.

2.2.1 Steps-1: Data Collection

Collection of the quality data relevant to the problem.

2.2.2 Steps-2: Data Preparation

Accurate data format is essential to apply machine learning algorithms. Sometimes, collected dataset is found incomplete and inadequate. Data preparation is a machine learning tool used to gather, combine, clean, complete, transform and feature selection of raw data in well shape to make true prediction [14, 26, 27].

2.2.3 Steps-3: Choosing of Machine Learning Model

There are three types of machine learning algorithms, i.e., supervised machine learning, unsupervised machine learning and reinforcement machine learning [5, 28]. The following condition is applied to choose the ML model. When class label (predicted variable) is categorical, the well-known algorithms applied as decision tree, random forest, support vector machines and artificial

neural networks, etc. When class label (predicted variable) is real values or continuous, the well-known algorithms applied are linear regression, decision tree regression, random forest regression and support vector regression, etc.

2.2.4 Steps-4: Mutually Exclusive Randomized Data Partition

The whole dataset *(d)* is split into two mutually exclusive randomized datasets called training and testing datasets [29, 30]. The training dataset (d_{train}) used to train the ML model. For the current study 75% of datasets is used to train the model. The remaining 25% used as test dataset ($d_{test} = (d) - d_{train}$) and used to evaluate the ML model performance for unseen datasets.

2.2.5 Steps-5: ML Model Evaluation, Model Complexity, Over and Under-Fit Model

Evaluation of ML model is assessed for training and testing algorithm. The ML model error found in training datasets called bias or in sample error or training error. The ML model error found in testing dataset called testing error or out-of-sample error or variance. The model with lowest values of root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE) and highest performance score (R^2), viewed best model. A model depicts low bias and high variance called over fit, while a model depicts high bias and high variance called under fit. Bias and variance tradeoff is common property of ML model [31, 32]. Optimum model complexity means a condition, where model depict optimum lowest range of bias and variance (Figure 1). The following equation is used to determine the prediction error (PE).

$$y = h(x) + e \quad (1)$$

Where *"e"* is error term. Let the estimate of *"h(x)"* is *"$h^*(x)$"* and the expected prediction squared error at *"x"* is determine as under:

$$\mathrm{PE}(x) = \mathrm{E}[(y - h^*(x))^2] \quad (2)$$

The prediction error further decomposed into bias and variance components as:

$$\mathrm{P.E}(x) = \mathrm{Var}[h^*(x)] + [\mathrm{Bias}\{h^*(x)\}]^2 + \mathrm{Var}(\mathrm{error}) \quad (3)$$

$$\mathrm{P.E}(x) = \text{variance} + \text{bias}^2 + \text{error} \quad (4)$$

2.2.6 Step-6: Hyper Parametric Tuning of ML Model

The ML models involve different parameters learning rates, etc. The process to extract the optimum values of these parameters learning rates called hyper-parameters tuning of ML problem. Hyper parameter tuning is a way to produce optimized ML model to avoid under and over fit conditions [33, 34]. The prominent ML function called GridSearchCV is used to tune the ML model.

2.2.7 Step-7: Model Deployment for Final Predictions

ML algorithms are applied after data preparation and hyper parametric tuning, and used to test/unseen data with the aim to make the true prediction for the dependent variable (wheat productivity). For the current study the following ML models are applied.

- Multiple linear regression
- Decision tree regression
- Random forest regression
- Support vector regression

2.3 Multiple Linear Regression Model (MLR)

The entire dataset is split into two mutually exclusive datasets, i.e., training and testing datasets. Multiple linear regression model is deployed for the dependent variable (wheat productivity) using the features (independents variables).

$$y_i = \beta_i z_i + \varepsilon \quad (5)$$

Where *"y_i"* stands wheat productivity in mds/ acre, *"z_i"* stands for the features, *"β_i"* stands for regression coefficients of features.

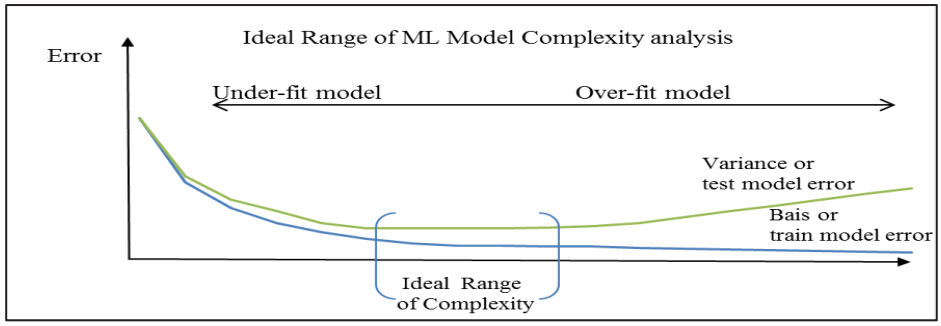

Fig. 1. Ideal range of model complexity for ML model

2.4 Decision Tree Regression Algorithm (DTR)

Decision tree is supervised machine learning algorithms used to breaks down the dataset into smaller homogenous subset using flowchart structure of root-to-leaf direction [35, 36]. Figure 2 shows the root-to-leaf direction flowchart of decision tree algorithms. The each homogenous subset is associated with specific decision about the whole data. The roots node is split into intermediate nods based on specifics characteristic of root node, and signifies a test. The intermediate node is further split into more intermediate or leaf nodes based on specifics characteristic of previous intermediate nodes, and each leaf node signifies the decision or outcomes. For the current study, the root nodes are constructed for the wheat yield using the different agronomical constrains. The mean square error (MSE) is used to build the structure of root-to-leaf nodes.

2.5 Random Forest Regression

Random forest regression (RFR) is supervised machine learning algorithm that combined the large set of decision trees [37-39]. RFR search has the best features and uses precision techniques to make the forest random. Figure 3 shows the flowchart structure of RFR algorithm. For RFR, decision tree is constructed for a random set of variable dataset. In RFR, various trees are merged together by taking the average of the prediction trees. RFR algorithm is based on three hyper parameters (no. of trees forest, no. of features and maximum depth of the tree). Machine learning function named GridSearchCV is used to tune the RFR hyper parameters.

2.6 Support Vector Regression (SVR)

Support vector is a non-linear generalization of generalized portrait algorithm developed by Vladimir Vapnik and his colleagues [40, 41]. SVR is based on the structural risk minimization principle for minimizing a bound on the generalization prediction error and to fit the optimum hyper plane with the aim to have maximum number of points within the decision boundary line. SVRs are limiting the prediction error margin (ε) and search to fit maximum points between the decision boundary

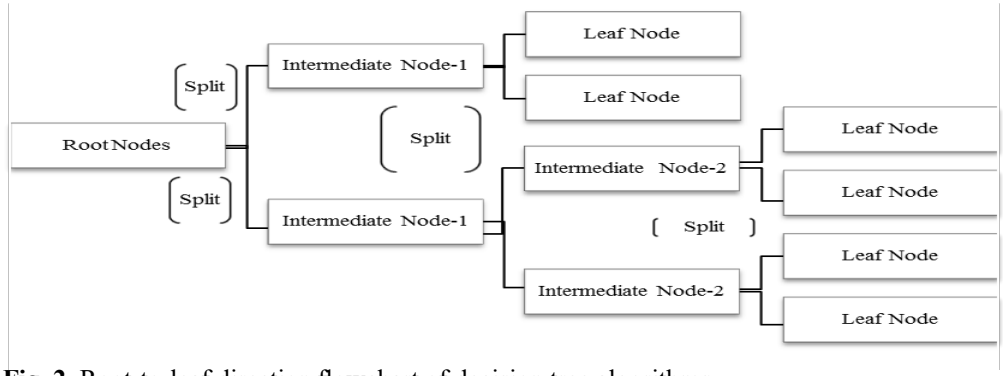

Fig. 2. Root-to-leaf direction flowchart of decision tree algorithms

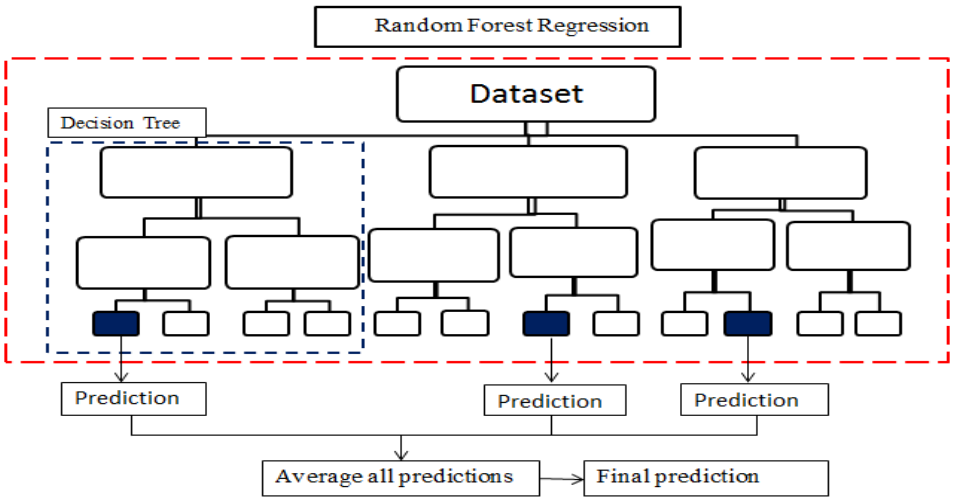

Fig. 3. Flowchart structure of RFR algorithm

lines. The SVRs are dependent on different kernel function such as linear, radial basis function (RBF) and polynomial. RBF kernel can better fit and handle the situation in case of nonlinear relation. Smola and Scholkopf [41] defined for a given train set $[(x_1, y_1),\ldots,(x_\ell, y_\ell)] \subset X \times R$, where *"X"* is the space of the input patterns $(X = R^d)$. The basic aim of this technique is to obtain *"f(x)"* has at most one *"ε"* deviation obtained from the current targets $"y_i"$ at the time and that is as flat as possible. The flatness is determine by the small value of *"w"* and the problem can be written as:

$$f(x) = (w, x) + b \ \ with \ \ w \epsilon x, b \epsilon R \qquad (6)$$

A linear function that approximate all pairs (x_i, y_i) with precision (ε).

$$\text{Min } \frac{1}{2}||w||^2 \qquad (7)$$

$$\text{Subject to } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \end{cases}$$

Figure 4 and Figure 5 show the structure for SVR. The key hyper parameters in SVR are noted as:

- ***Hyperplane****:* Hyper plane means a line used to separate two datasets
- ***Support vector***: Datasets points lies on either side of the hyper plane as closest as, called support vectors.
- ***Kernel****:* Kernel means a set of mathematical functions used to transform input data into required forms and it is generally used to search optimum hyper plane.
- ***Decision boundary lines:*** The two lines drawn around the hyper plane at a distance of *" "* and used to create a margin between the dataset points.

2.7 K-Fold Cross-Validations Measures

Cross-validation is resampling, out of sample testing or rotation estimations, that partitioning the data into complementary subsets. In K-Fold cross validations various rounds of train test split are applied and average of all rounds of partitions is taken to assess the model performance [42-44]. For K-Fold cross- validation, various learning disjoint subsets of equal size *"k"* is made and model is trained for *"k-1"* subsets, and test for complementary subset. For the current study "10" fold cross- validation is applied. Figure 6 shows the flowchart for the application of K-Fold cross-validation.

3. RESULTS AND DISCUSSION

3.1 Randomized Data Partitions

For the current study collected datasets is complete and free from any inadequacy. Data partition is applied using the randomized data splitting. The 75% datasets is used to train the model and 25% is used to test the model. The centroid clustering scheme is applied to prepare the dataset of 136 tehsils zones in Punjab for wheat crop CCE. The wheat crop CCE of 102 tehsils is split to train the

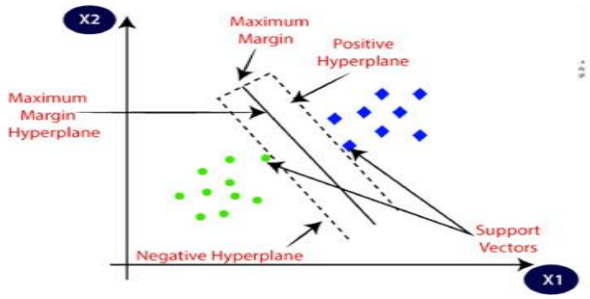

Fig. 4. Structure of hyper plane for SVR

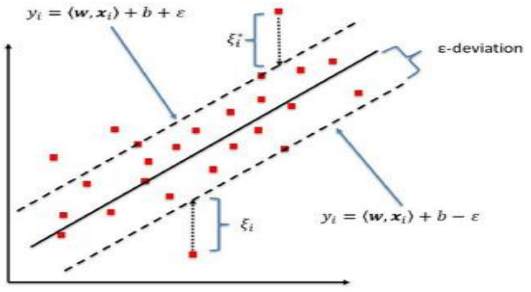

Fig. 5. Prediction error margin of SVR

K-Fold cross-validation using randomized data partitions							
Average of Fold	Fold-1	**Test**	Train	Train	Train	Train	Test score -1
	Fold-2	Train	**Test**	Train	Train	Train	Test score -2
	Fold-3	Train	Train	**Test**	Train	Train	Test score -3
	Fold-4	Train	Train	Train	**Test**	Train	Test score -4
	Fold-5	Train	Train	Train	Train	**Test**	Test score -5

Fig. 6. Flowchart structure of K-Fold cross-validation

model, and 34 tehsils is split to test the model performance for unseen datasets.

3.2 Hyper Parametric Tuning of Machine Learning Models

Figure 7 shows the machine learning model complexity for DTR and RFR. Learning curves show high variance and high bias at lower model complexity (under fit). The learning curves reached their lowest point at the ideal range of model complexity with an optimum depth value of 4. However, the test curve rises again (over fitting) after reaching the lowest point in the optimum range of model complexity. GridSearchCV, a Python's key hyper parameter tuning library, produced the optimal value for the minimum sample split is 4, while the number of trees in the forest is found optimum at 10 for RFR.

3.3 Integrating the Performance of MLR, DTR, RFR and SVR

Table 1 shows the performance of multiple regression, decision tree regression, random forest regression and support vector regression using the machine learning approach. The highest value of R^2 found is 0.90 for RFR, followed by DTR, MLR and SVR for train model ($R^2_{RFR} > R^2_{DTR} > R^2_{MLR} > R^2_{SVR}$), while for the test model, it is 0.74, followed by MLR, DTR and SVR ($R^2_{RFR} > R^2_{MLR} > R^2_{DTR} > R^2_{SVR}$). Comparing the performance of MSE and MAE, the lowest MSE and MAE are reported for the RFR as 4.58 and 1.66, followed by DTR, MLR and SVR for the train model ($MSE\ (MAE)_{RFR} < MSE\ (MAE)_{DTR} < MSE\ (MAE)_{MLR} < MSE\ (MAE)_{SVR}$). In the test model, RFR produced an MSE of 8.98 and an MAE of 2.36, followed by MLR, DTR, and SVR ($MSE\ (MAE)_{RFR} < MSE\ (MAE)_{MLR} < MSE\ (MAE)_{DTR} < MSE\ (MAE)_{SVR}$). Integrating the performance of prediction error (P.E) for the MSE and MAE, lowest value of P.E found for the RFR is 13.56 for MSE and 4.02 for MAE, and followed by DTR, MLR and SVR. The relationship among the prediction error (P.E) values for RFR, DTR, MLR, and SVR is represented as follows $P.E_{RFR} < P.E_{DTR} < P.E_{MLR} < P.E_{SVR}$. Figure 8 shows the learning curves for the MSE and Figure 9 shows the learning curves for the MAE. It is found that RFR is better and optimized fitted model, as the lowest error is found for the RFR with highest performance score (R^2).

3.4 K-fold Cross-Validations Measures

Table 2 shows the K-Fold cross-validation measures for the MLR, DTR, RFR and SVR using the MAE and MSE criterion. For the MLR, the smallest values of MAE and MSE are found for K-Fold-5, with 2.25 and 7.65, respectively, while the largest values are found for K-Fold-7, with 3.74 for MAE

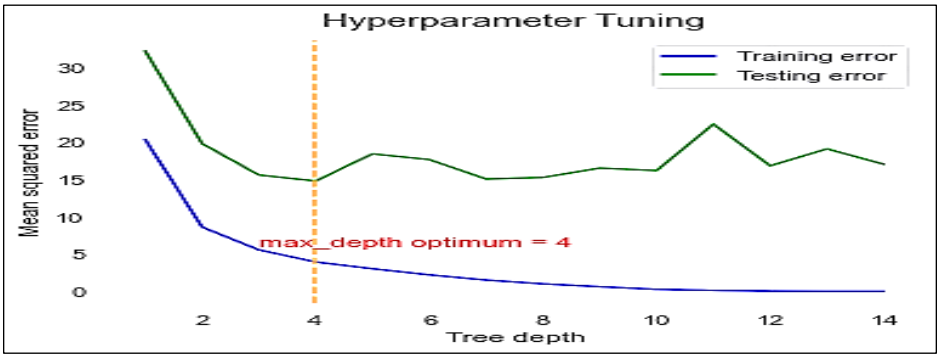

Fig. 7. Machine learning model complexity for DTR and RFR

Table 1. Performance of machine learning models

Evaluation Metrics	Train set				Test set			
	MLR	DTR	RFR	SVR	MLR	DTR	RFR	SVR
R^2	0.75	0.89	0.90	0.48	0.68	0.508	0.74	0.31
MSE	11.09	4.65	4.58	22.9	11.23	17.36	8.98	24.17
MAE	2.76	1.67	1.66	3.55	2.59	3.06	2.36	3.67
P.E (MSE)	22.32	22.01	13.56	47.07	--	--	--	--
P.E (MAE)	5.35	4.72	4.02	7.22	--	--	--	--

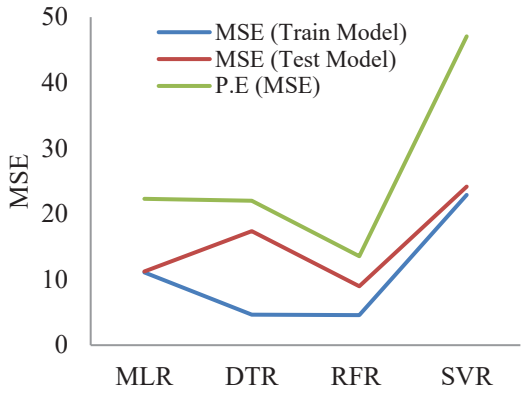

Fig. 8. Learning curves for the MSE

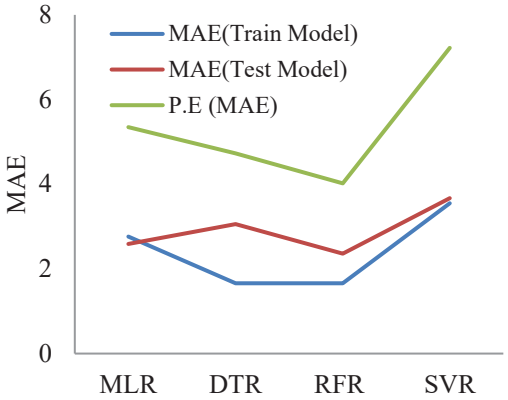

Fig. 9. Learning curves for the MAE

Table 2. K-Fold cross-validation measures for the MLR, DTR, RFR and SVR

ML models	Criterions	K-Fold average	K-Fold 1	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K - Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10
MLR	MAE	2.97	2.65	3.22	2.30	2.45	2.25	3.14	3.74	3.16	3.10	3.74
	MSE	13.10	9.27	16.04	8.19	10.15	7.65	15.78	19.96	13.29	14.59	16.71
DTR	MAE	2.50	2.23	2.02	2.11	1.90	2.86	3.20	3.17	2.78	2.06	2.08
	MSE	10.56	7.55	12.12	7.38	6.96	12.05	16.16	15.97	9.96	7.82	9.00
RFR	MAE	2.29	2.08	2.24	2.18	1.55	1.99	2.24	3.49	2.79	2.04	2.45
	MSE	8.41	5.84	6.68	6.20	4.50	7.53	6.92	20.04	9.83	8.16	10.77
SVR	MAE	3.57	2.93	2.94	3.04	2.36	3.62	3.43	4.42	5.08	3.69	4.18
	MSE	21.80	14.55	12.72	20.93	7.17	21.73	21.22	29.26	35.69	21.89	32.93

and 19.96 for MSE. The lowest values of MAE and MSE are found at K-Fold-4, with 1.90 and 6.96 for DTR, 1.55 and 4.50 for RFR, and 2.36 and 7.17 for SVR, respectively. The largest values of MAE and MSE are found to be 3.20 and 16.16 for the DTR at K-Fold-6, 3.49 and 20.04 for RFR at K-Fold-7, and 5.08 and 35.69 for SVR at K-Fold-8. Figure 10 shows the learning curves of MAE and MSE for comparing RFR, DTR, MLR and SVR. RFR are found to perform better as it predicted the lowest learning curve compared to other machine learning models.

4. CONCLUSIONS

Machine learning (ML) emerged as an advanced tools of data science and artificial intelligence. ML is used to extract reliable information from data using various algorithms. The issue of precise crop prediction gained worldwide attention in the midst of food security concerns. ML models, i.e., linear regression, decision tree, random forest and support vector regression are applied as an alternative of statistical model in almost every field. True crop prediction got key significance in the midst of

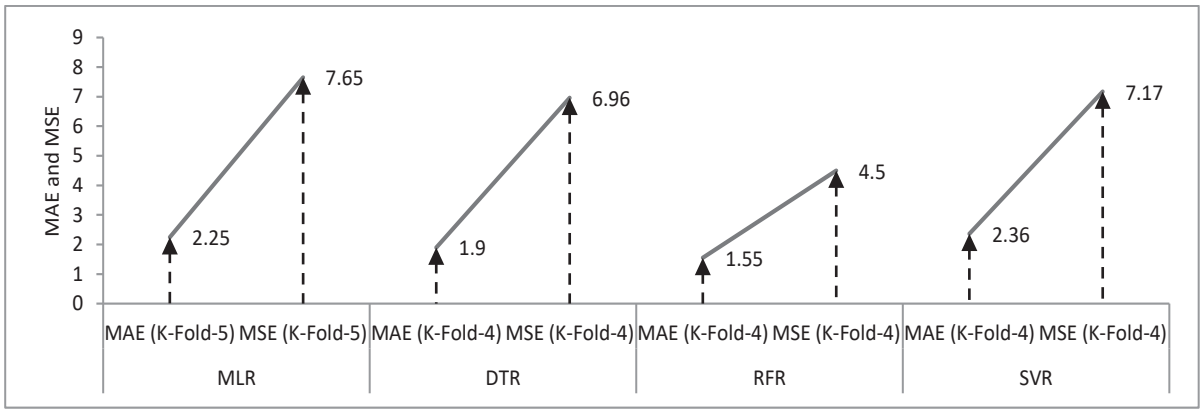

Fig. 10. Learning curves of MAE and MSE for K-Fold cross-validations measures

food security situation in the world. Validations and optimization of ML algorithms have become the key concerns for the world to predict the true agricultural productions. This study integrated the efficacies of different machine learning algorithms, i.e., multiple linear regression (MLR), decision tree regression (DTR), random forest regression (RFR), and support vector regression (SVR) with the aim to get optimized model to predict the wheat productivity in Pakistan using various agronomical constrains. The updated centroid clustering based dataset is collected from the Crop Reporting Service. The python's key library called ScikitLearn is used to analyze the results. Different machine learning tools such as randomized data partitions, hyper parametric tuning, complexity analysis and cross-validation measures are also performed to get the optimized model.

Highest value of R^2 is reported from RFR, followed by DTR, MLR and SVR for train model ($R^2_{RFR} > R^2_{DTR} > R^2_{MLR} > R^2_{SVR}$), and for test model, it is followed by MLR, DTR and SVR ($R^2_{RFR} > R^2_{MLR} > R^2_{DTR} > R^2_{SVR}$). Lowest MSE and MAE are reported for the RFR, followed by DTR, MLR and SVR for the train model ($MSE\ (MAE)_{RFR} < MSE\ (MAE)_{DTR} < MSE\ (MAE)_{MLR} < MSE\ (MAE)_{SVR}$), and for the test model, it is followed by MLR, DTR and SVR ($MSE\ (MAE)_{RFR} < MSE\ (MAE)_{MLR} < MSE\ (MAE)_{DTR} < MSE\ (MAE)_{SVR}$). Lowest value of prediction error (P.E) is found from RFR, followed by DTR, MLR and SVR both for the MSE and MAE ($P.E_{RFR} < P.E_{DTR} < P.E_{MLR} < P.E_{SVR}$). Learning curves of MAE and MSE depict that RFR is an optimized machine learning model compared to DTR, MLR, and SVR. K-Fold cross-validation results indicate that the lowest error is found for RFR when compared with DTR, MLR, and SVR. RFR found a validated and optimized model when compared with other machine learning models. This research can also be extended by applying other clustering approaches and machine learning algorithms.

5. CONFLICT OF INTEREST

The authors declare no conflict of interest.

6. ACKNOWLEDGEMENT

All authors acknowledge the strong data collection mechanism of Crop Reporting Service, Punjab being a sole source of agricultural crop statistics in Punjab, Pakistan.

7. REFERENCES

1. J.A. Polonsky, A. Baidjoe, Z.N. Kamvar, A. Cori, K. Durski, W.J. Edmunds, R.M. Eggo, S. Funk, L. Kaiser, P. Keating, O.P. Waroux, M. Marks, P. Moraga, O. Morgan, P. Nouvellet, R. Ratnayake, C.H. Roberts, J. Whitworth and T. Jombart. Outbreak analytics: a developing data science for informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B* 374: 1-11 (2019).
2. H. Yang. Building an Agile and Scalable Data Science Organization; (Chapter 3) In: Data Science, AI, and Machine Learning in Drug Development. *Chapman and Hall/CRC* (2022).
3. T.H. Davenport. From analytics to artificial intelligence. *Journal of Business Analytics* 1(2): 73-80 (2018).
4. Y, Nazarathy, and H. Klok. Statistics with Julia: Fundamentals for data science, machine learning and artificial intelligence. *Springer Nature* (2021).
5. B. Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)* 9: 381-386 (2020).
6. C. Park, C.C. Took, and J.K. Seong. Machine learning in biomedical engineering. *Biomedical Engineering Letters* 8: 1-3 (2018).
7. N. Yadav, S.M. Alfayeed, and A. Wadhawan. Machine learning in agriculture: techniques and applications. *International Journal of Engineering Applied Sciences and Technology* 5(7): 118-122 (2020).
8. K. Patel, and H.B. Patel. A comparative analysis of supervised machine learning algorithm for agriculture crop prediction. *Paper presented at the 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (2021).
9. A.M. Lad, K.M. Bharathi, B.A. Saravanan, and R. Karthik. Factors affecting agriculture and estimation of crop yield using supervised learning algorithms. *Materials Today: Proceedings* 62: 4629-4634 (2022).
10. K.G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis. Machine learning in agriculture: A review. *Sensors* 18(8): 1-29 (2018).
11. Y. Mekonnen, S. Namuduri, L. Burton, A. Sarwat, and S. Bhansali. Machine learning techniques in wireless sensor network based precision agriculture. *Journal of the Electrochemical Society* 167(3): 1-10 (2019).
12. E. Kamir, F. Waldner, and Z. Hochman. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing* 160:124-135 (2020).
13. M.S. Ud-Din, M. Mubeen, S. Hussain, A. Ahmad, N. Hussain, M.A. Ali, A.E. Sabagh, M. Elsabagh, G.M Shah, S.A. Qaisrani, M. Tahir, H.M.R. Javeed, M.A. H. Ali and W. Nasim. World nations priorities on climate change and food security. *Springer* pp. 365-384 (2022).
14. M. Islam, and F. Shehzad. A prediction model optimization critiques through centroid clustering by reducing the sample size, integrating statistical and machine learning techniques for wheat productivity. *Scientifica* 1-11 (2022).

15. D. Elavarasan, and P.D.R. Vincent. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. *Journal of Ambient Intelligence and Humanized Computing* 12: 10009–10022 (2021).
16. M. Islam, F. Shehzad, A. Qayyum, M.W. Abbas, and R. Siddiqui. Growth analysis of production of food crops and population growth for Food Security in Pakistan. *Proceedings of the Pakistan Academy of Sciences: B. Life and Environmental Sciences* 60(1): 83-90 (2023).
17. P. Feng, B. Wang, D.L. Liu, C. Waters, and Q. Yu. Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. *Agricultural and Forest Meteorology* 275: 100-113 (2019).
18. O. Erenstein, M. Jaleta, K. Sonder, K. Mottaleb, and B. Prasanna. Global maize production, consumption and trade: trends and R&D implications. *Food Security* 14(5): 1295-1319 (2022).
19. P. Giraldo, E. Benavente, F. Manzano-Agugliaro, and E. Gimenez. Worldwide research trends on wheat and barley: A bibliometric comparative analysis. *Agronomy* 9(7): 1-18 (2019).
20. R. Ghimire, H. Wen-Chi, and R.B. Shrestha. Factors affecting adoption of improved rice varieties among rural farm households in Central Nepal. *Rice Science* 22(1): 35-43 (2015).
21. S. Ali, Y. Liu, M. Ishaq, T. Shah, A. Ilyas, and I.U. Din. Climate change and its impact on the yield of major food crops: Evidence from Pakistan. *Foods* 6(6): 1-19 (2017).
22. A. Chlingaryan, S. Sukkarieh, and B. Whelan. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture* 151: 61-69 (2018).
23. D.A. Bondre, and S. Mahagaonkar. Prediction of crop yield and fertilizer recommendation using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology* 4(5): 371-376 (2019).
24. L.S. Cedric, W.Y.H. Adoni, R. Aworka, J.T. Zoueu, F.K. Mutombo, M. Krichen, and C.L.M. Kimpolo. Crops yield prediction based on machine learning models: case of west African countries. *Smart Agricultural Technology* 2: 1-14 (2022).
25. S. Mohapatra, and N. Chaudhary. Statistical analysis and evaluation of feature selection techniques and implementing machine learning algorithms to predict the crop yield using accuracy metrics. *Engineered Science* 21: 1-11 (2022).
26. V. Gudivada, A. Apon, and J. Ding. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* 10(1): 1-20 (2017).
27. A. Dogan, and D. Birant. Machine learning and data mining in manufacturing. *Expert Systems with Applications* 166(2): 114060 (2021).
28. M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A.J Aljaaf. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In: Supervised and Unsupervised Learning for Data Science. *Springer* pp. 3-21 (2020).
29. J.J. Salazar, L. Garland, J. Ochoa, and M.J. Pyrcz. Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. *Journal of Petroleum Science and Engineering* 209: 109885 (2022).
30. J. Tan, J. Yang, S. Wu, G. Chen, and J. Zhao, A critical look at the current train/test split in machine learning. *arXiv preprint arXiv*:2106-04525 (2021).
31. M. Islam. Integrating Statistical and Machine Learning Techniques to Predict Wheat Production in Pakistan. Ph.D. Thesis, Department of Statistics, *The Islamia University of Bahawalpur, Pakistan* (2022).
32. P. Chakraborty, S.S. Rafiammal, C. Tharini, and D.N. Jamal. Influence of bias and variance in selection of machine learning classifiers for biomedical applications smart data intelligence. *Proceedings of ICSMDI, Springer:* 459-472 (2022).
33. A. Vabalas, E. Gowen, E. Poliakoff, and A.J. Casson. Machine learning algorithm validation with a limited sample size. *PLoS One* 14(11): e0224365 (2019).
34. L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-Tzur, M. Hardt, B. Recht and A Talwalkar. A system for massively parallel hyper parameter tuning. *Proceedings of Machine Learning and Systems* 2: 230-246 (2020).
35. F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)* 48(3): 128-138 (2017).
36. B.T. Jijo, and A.M. Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends* 2(1): 20-28 (2021).
37. M. Maduranga, and R. Abeysekera. Treeloc: An ensemble learning-based approach for range based indoor localization. *International Journal of Wireless and Microwave Technologies (IJWMT)* 11(5): 18-25 (2021).
38. K. Rawal, and A. Ahmad. A comparative analysis of supervised machine learning algorithms for electricity demand forecasting. *Paper presented at the 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T):* (2022).
39. F. Farooq, A.M. Nasir-Amin, K. Khan, M. Rehan-Sadiq, M.F. Javed, F. Aslam, and R. Alyousef. A comparative study of random forest and genetic engineering programming for the prediction of compressive strength of high strength concrete (HSC). *Applied Sciences* 10(20): 7330 (2020).
40. H. Drucker, C.J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems* 9: 155-161 (1997).
41. A.J. Smola, and B. Scholkopf. A tutorial on support vector regression. *Statistics and Computing* 14(3):

199-222 (2004).
42. R.T. Nakatsu. An evaluation of four resampling methods used in machine learning classification. *IEEE Intelligent Systems* 36(3): 51-57 (2020).
43. B. Bischl, O. Mersmann, H. Trautmann, and C. Weihs. Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation* 20(2): 249-275 (2012).
44. G. Afendras, and M, Markatou. Optimality of training/test size and resampling effectiveness in cross-validation. *Journal of Statistical Planning and Inference* 199: 286-301 (2019).