Advances in Research

21(9): 98-105, 2020; Article no.AIR.60327
ISSN: 2348-0394, NLM ID: 101666096

Assessment of the Different Machine Learning Models for Prediction of Cluster Bean (*Cyamopsis tetragonoloba L. Taub.*) Yield

Darshan Jagannath Pangarkar[1*], Rajesh Sharma[2], Amita Sharma[3] and Madhu Sharma[2]

[1]College of Agriculture (COA), Swami Keshwanand Rajasthan Agriculture University (SKRAU), Bikaner, India.
[2]Department of Agricultural Economics, College of Agriculture (COA), Swami Keshwanand Rajasthan Agriculture University (SKRAU), Bikaner, India.
[3]Institute of Agri Business Management (IABM), Swami Keshwanand Rajasthan Agriculture University (SKRAU), Bikaner, India.

Authors' contributions

This work was carried out in collaboration among all authors. Author DJP worked on framing the research design, applications of different models of machine learning for prediction of yield and writing of the research paper. Author RS supervised the research work. He conceptualized the exploration of different machine learning models for yield prediction and supported for analysis. Author AS supported for application of machine learning and analysis of data for research purpose. Author MS helped for overall implementation and research reviews writing. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AIR/2020/v21i930238
Editor(s):
(1) Prof. Magdalena Valsikova, Slovak University of Agriculture, Slovakia.
Reviewers:
(1) Junaidi Sungei Putih, Indonesian Rubber Research Institute, Indonesia.
(2) Jose Antonio Valles, Mexican Institute for Logistics and Supply Chain, Mexico.
Complete Peer review History: http://www.sdiarticle4.com/review-history/60327

Short Research Article

Received 12 June 2020
Accepted 18 August 2020
Published 27 August 2020

ABSTRACT

Prediction of crop yield can help traders, agri-business and government agencies to plan their activities accordingly. It can help government agencies to manage situations like over or under production. Traditionally statistical and crop simulation methods are used for this purpose. Machine learning models can be great deal of help. Aim of present study is to assess the predictive ability of various machine learning models for Cluster bean (*Cyamopsis tetragonoloba L. Taub.*) yield

**Corresponding author: E-mail: pangarkar.dj.ae@gmail.com;*

prediction. Various machine learning models were applied and tested on panel data of 19 years i.e. from 1999-2000 to 2017-18 for the Bikaner district of Rajasthan. Various data mining steps were performed before building a model. K- Nearest Nighbors (K-NN), Support Vector Regression (SVR) with various kernels, and Random forest regression were applied. Cross validation was also performed to know extra sampler validity. The best fitted model was chosen based cross validation scores and R2 values. Besides the coefficient of determination (R2), root mean squared error (RMSE), mean absolute error (MAE), and root relative squared error (RRSE) were calculated for the testing set. Support vector regression with linear kernel has the lowest RMSE (23.19), RRSE (0.14), MAE (19.27) values followed by random forest regression and second-degree polynomial support vector regression with the value of gamma = auto. Instead there was a little difference with R^2, placing support vector regression first (98.31%), followed by second-degree polynomial support vector regression with value of gamma = auto (89.83%) and second-degree polynomial support vector regression with value of gamma = scale (88.83%). On two-fold cross validation, support vector regression with a linear kernel had the highest cross validation score explaining 71% (+/- 0.03) followed by second-degree polynomial support vector regression with a value of gamma = auto and random forest regression. KNN and support vector regression with radial basis function as a kernel function had negative cross validation scores. Support vector regression with linear kernel was found to be the best-fitted model for predicting the yield as it had higher sample validity (98.31%) and global validity (71%).

Keywords: Yield; machine learning; K-NN; SVR; random forest.

1. INTRODUCTION

Improvement in Information technology has permitted digitalization in every sector of the economy. Agriculture is not an exception to it. New concepts of Machine Learning and Artificial Intelligence have the potential to revolutionize the way in which we collect and analyse agricultural data. Machine learning is a branch of Artificial Intelligence, it gives computers the ability to learn without being explicitly programmed. 'Integrating computer science with agriculture helps in forecasting crops' [1]. Crop yield predictions can help the government in the formulation of suitable policies regarding imports-exports, procurement, and in managing situations like over-production and under-production. This estimation can also help traders and agribusinesses.

This study is intended to predict Cluster bean yield. Cluster bean is most important commercial crop of arid and semi-arid region with average yield of 500-700 kg/ha in India. Seed of cluster bean contains 30-33% gum in the endosperm. Guar gum is an important industrial product. India produces about 80% of worlds cluster bean and about 75% of guar gum is exported.

Linear regression is commonly used for crop yield prediction but has weak results [2]. Machine learning techniques are based on non-parametric and semi-parametric structures, cluster with validation rely on predictive accuracy [3]. Regression trees and Random forest [4], KNN [5], and support vector regression [6] are common machine learning techniques applied for the purpose. Some comparisons have been made, looking for the most accurate technique. Drummond et al. 2003 studied site specific yield prediction with statistical and neural methods for soybean and corn [7]. Gonzalez-Sanchez et al. 2014 compared predictive ability several machine learning methods and found the M5-prime decision tree and KNN regression as best fits [2]. Devika and Ananthi, 2018 compared the accuracy of KNN and linear regression techniques [8]. Machine learning presents several methods to define rules and patterns in large data sets related to crop yield and has well known to predict capability [9].

2. MATERIALS AND METHODS

Cluster bean (*Cyamopsis tetragonoloba L. Taub*.) crop was selected for study as it is an important commercial crop of Rajasthan. Average yield of guar in Rajasthan is 319 kg/ha. Bikaner district was selected for study as it has the highest area (7.09 lakh ha) under Cluster bean cultivation during 2017-18 (Rajasthan Agricultural Statistics at a Glance 2018). The study was based on secondary data collected from various published sources like Rajasthan Agricultural Statistics at a Glance, indiastat.com, data.gov.in, etc. Panel data of various independent variables (Table 1) and dependent variable (yield in kg/ha) were collected from the

year 1999-2000 to 2017-18. In this work, experiments were performed using python based SpyderIDE. Collected data were screened for missing values. Independent variables were normalized wherever needed with suitable transformation. Data were standardized using *StandardScaler()* python library. 80% of data were used to train models and 20% of data were used to test the trained models. In this work K-Nearest Nighbors regression (KNN), Support Vector Regression (SVR) with Linear, Polynomial, and Radial Basis Function kernel and Random Forest regression were used and tested.

2.1 K-Nearest Nighbors

K-Nearest Nighbors makes predictions based on *K* nighbors closest to that point. KNN predictions are based on the assumption that objects close in distance are partially similar. Euclidean distance between points was calculated using the formula:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}$$

Where

x_i and y_i are query points and case of example sample, respectively.

A new point was predicted based on the mean of *K* labels.

The KNN technique has been used to study the behaviour of the crop [10]. Nevertheless, very few comparisons of KNN against other machine learning methods applied to CYP have been made. This work applies a KNN algorithm to predict the Cluster bean yield, and the results were compared against SVR, and Random forest regression. K value of 5 and the Euclidean distance were used as parameters for this technique.

2.2 Support Vector Regression

The support vector machine was first introduced for classification later it was extended for regression problems and has been applied for a variety of problem statements. In this study, SVR was used with linear, polynomial, and radial basis function kernel. 'On its simplest form, the goal of the support vector technique is to obtain a linear function f(x) = <w,x> + b with $w \in R^N$ and $b \in R$ for a given training set $\{(x_1,y_1),...,(x_m,y_m)\}$. That function f(x) should have at most one ε deviation from the current obtained targets y_i at the time that is as flat as possible' [11]. Flatness can be obtained by a small value of w. Thus, the problem can be written as [12]:

$$\min \frac{1}{2}\|\omega\|^2 + c\sum_{i=1}^{m}(\xi_i + \xi_i^*)$$

$$Subject\ to \begin{cases} y_i - \langle w_1 x_i\rangle - b \le \varepsilon + \xi_i \\ \langle \omega_1 x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

Where ξ_i, ξ_i^* are slack variables and *C* is called the regulatory parameter and determines accepted deviation larger than ε. In most cases, these parameters can be easily estimated by using dual formula. Minimization dual formula for linear SVR:

$$L_{(\alpha)} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - a_i^*)(\alpha_j - \alpha_j^*)x_i' x_j + \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i)$$

Nonlinear SVR finds coefficients that minimize:

$$L_{(\alpha)} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - a_i^*)(\alpha_j - \alpha_j^*)Gx_i' x_j + \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i)$$

Where, $Gx_i' x_j$ is known as the Kernel Function, which allows projecting original data into higher dimensional space to be linearly separable.
Both subject to

$$\sum_{n=1}^{N}(\alpha_n - \alpha_n^*) = 0$$

$$\forall n: 0 \le x_n \le c$$

$$\forall n = 0 \le \alpha_n^* \le C$$

To obtain good predictions the parameters needed to be tuned. Unfortunately, there is no automated method to find such optimal values. Thus, these were established by the trial and error method.

Linear kernel function was expressed as $G(x'_n,x) = x'_n.x$.

Regression function for SVR with a linear kernel is expressed as

$$f(x) = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*)(x_n' x) + b$$

For nonlinear regression problems, the polynomial kernel was expressed as $G(x_j,x_k)=(1+x_j'x_k)^q$, where q is in the set {2,3,...}. and Radial basis function kernel was expressed

$$G(xj, xk) = \exp(-|xj - xk|2)$$

Regression function for nonlinear SVR can be expressed as

$$f(x) = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*) G(x_n' x) + b$$

Where, α_n, α_n^* are Langrange multipliers and x_n' are support vectors that are learned through optimization technique in SVM regression.

2.3 Random Forest Regression

Random forest regression is a modification over decision tree regression. It combines many decision trees into a single model. Random forest regression uses an ensemble learning method for regression. It operates by constructing a multitude of decision trees at training time and outputting the mean prediction of individual trees. A sample random forest regression tree is shown in Fig. 1. It limits the number of features that can be split on at each node. This ensures that the ensemble model doesn't rely heavily on any individual feature. Each tree draws a random sample from the original data set while generating splits that prevent overfitting. In this study number of trees, estimation was limited to 10 i.e. model has averaged the prediction of 10 decision trees to give a final estimate.

2.4 Accuracy Metrics

To compare sample validity of models' root mean squared error (RMSE), root relative squared error (RRSE), mean absolute error (MAE), and coefficient of determination (R^2) were estimated. To estimate the global validity of models' cross validation scores were estimated.

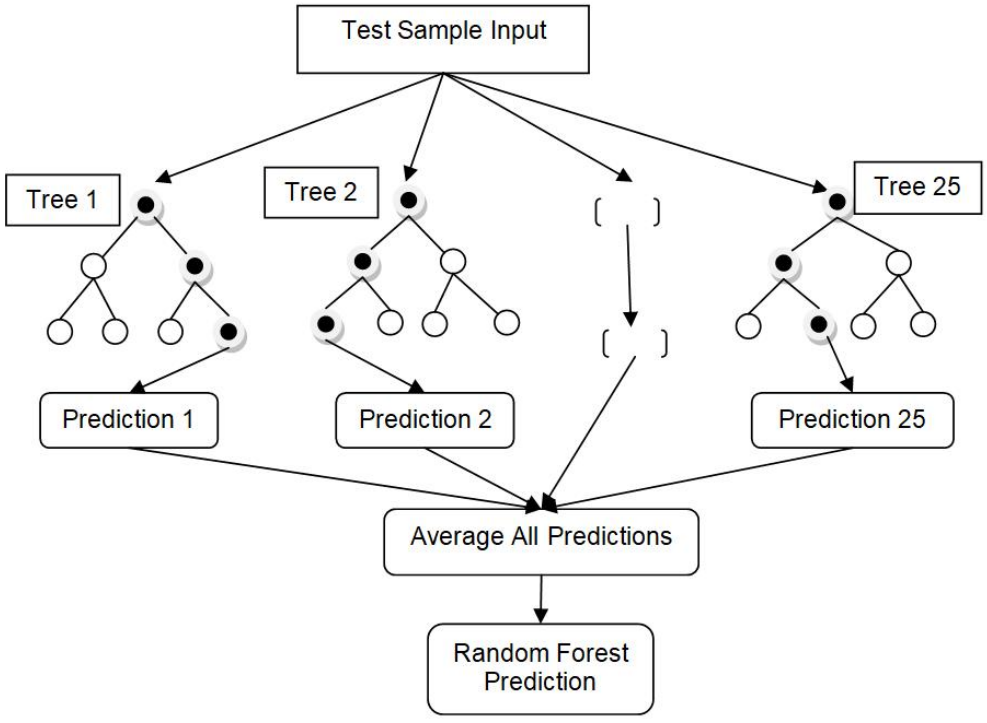

Fig. 1. Sample random forest regression tree

3. RESULTS AND DISCUSSION

Several evaluations of ML (Machine learning) methods have been applied to crop yield prediction in the literature, each with a different purpose. Some works measure ML performance using a particular attribute set [13], some have compared various ML methods [2,14,15]. However, this work is limited to one crop and its results are hard to extrapolate to other crops.

Table 1. Independent variables supplied to machine learning models

Variable		Measurement
X_1	Rainfall	mm
X_2	Seed distribution	qtl
X_3	Percent area under the plant protection (PAPP)	Percent
X_4	N	Tons
X_5	P	Tons
X_6	K	Tons
X_7	Previous year price	≠/qtl
X_8	Area under cultivation	ha
X_9	Area under irrigation	ha
X_{10}	Production	qtl

mm = millimeter; qtl = quintal; ha = hectares

Table 2. Performance metrics

RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$
RRSE	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
R^2	$1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert$

y_i = real value, y^_i = estimated value, i = observation, y¯ = mean
RMSE = root mean squared error, RRSE = root relative squared error
R^2 = coefficient of determination, MAE = mean absolute error

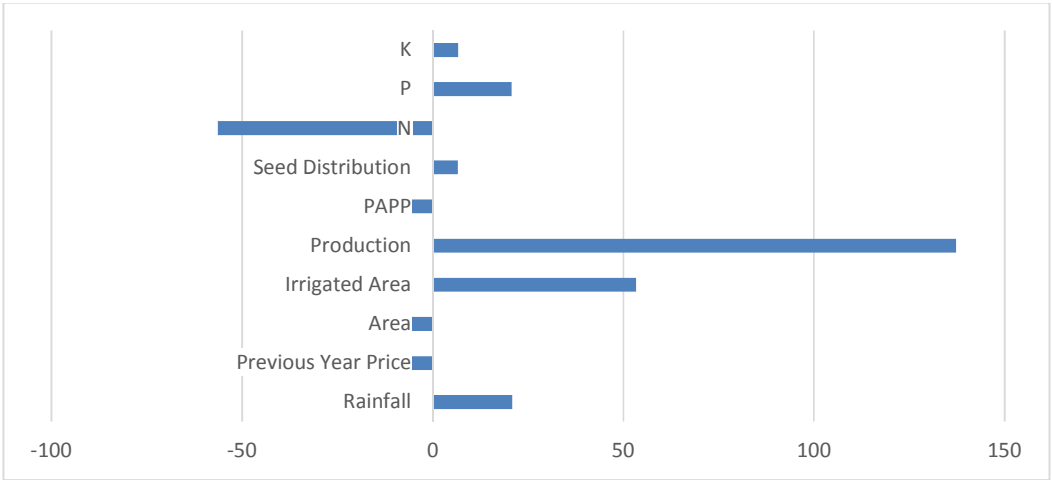

Fig. 2. Importance of explanatory variables in the support vector regression model with linear kernel

Table 3. Comparison of sample validity of models

Models		RMSE	RRSE	MAE	R^2 (%)
KNN		70.32	0.63	67.45	81.90
Linear SVR		23.19	0.14	19.27	98.31
Polynomial SVR	gamma = auto	53.94	0.41	51.96	89.35
	gamma = scale	55.24	0.42	52.99	88.83
RBF SVR	gamma= auto	89.87	0.66	77.98	70.44
	gamma = scale	94.02	0.69	81.80	67.65
RF Regression		63.30	0.53	62.65	85.33

KNN: K-Nearest Nighbors, SVR: Support vector regression, RBF: Radial Basis Function, RF regression: Random forest regression

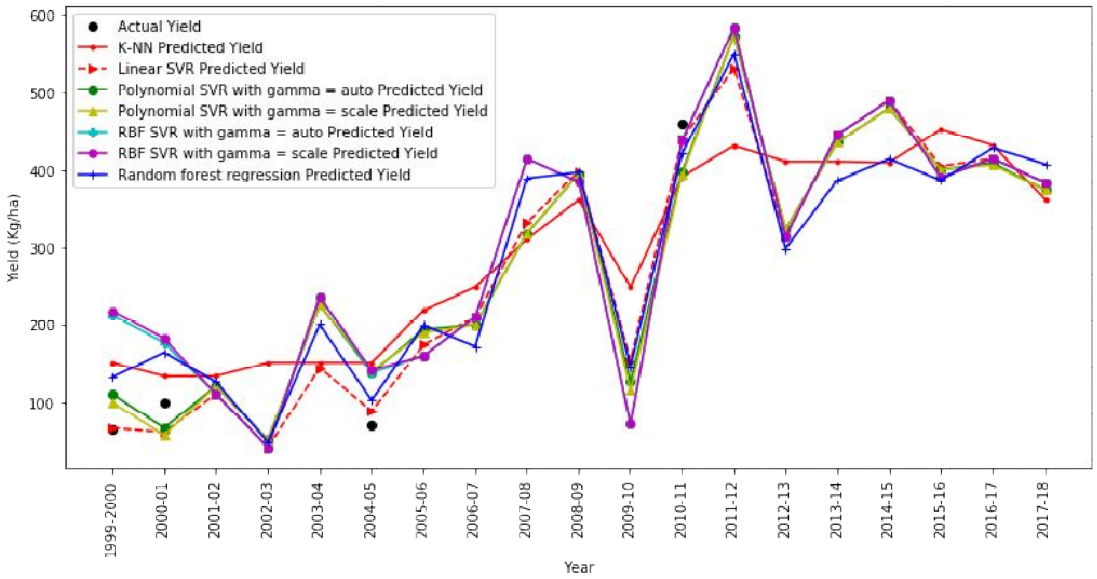

Fig. 3. Actual and predicted yield of sample validated models

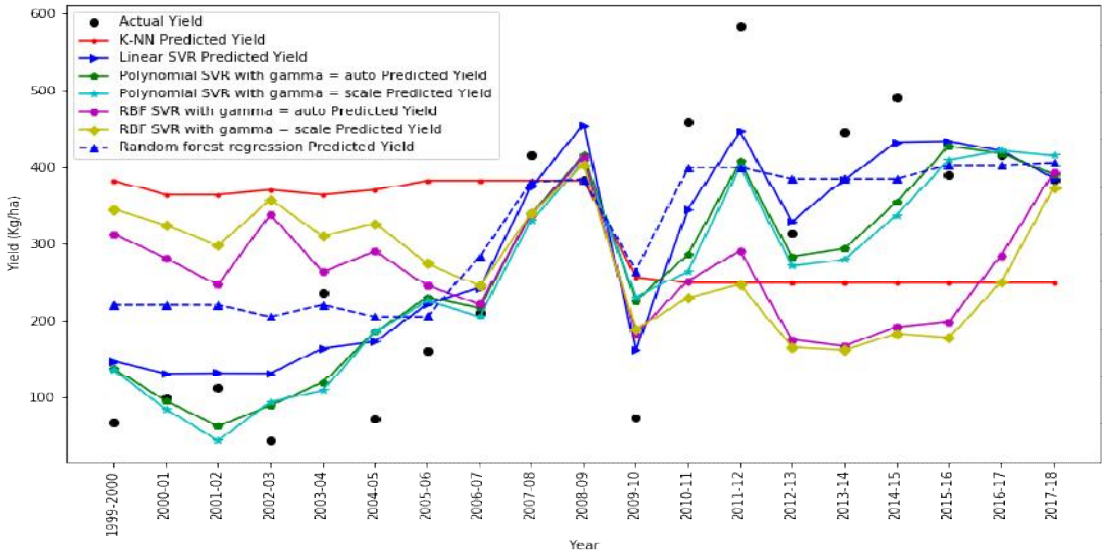

Fig. 4. Actual and predicted yield of cross validated models

Table 4. Comparison of global validity of models

Models		Two-fold cross validation scores
KNN		-1.80 (+/- 1.17)
Linear SVR		0.71 (+/- 0.03)
Polynomial SVR	gamma = auto	0.45 (+/- 0.49)
	gamma = scale	0.37 (+/- 0.59)
RBF SVR	gamma= auto	-1.01 (+/- 0.68)
	gamma = scale	-1.52 (+/- 0.51)
RF Regression		0.37 (+/- 0.01)

KNN: K-Nearest Nighbors, SVR: Support vector regression, RBF: Radial Basis Function, RF regression: Random forest regression, Values in parenthesis: mean standard error

Coefficients of the linear support vector regression model represent the relative importance of variables. Fig. 2 indicates the importance of explanatory variables for linear support vector regression. Production and irrigated area had the most important positive role while nitrogen consumption has the most important negative role in yield prediction. Irrigation has significant impact on guar yield [1].

Model comparison for sample validity was made based on four performance metrics. Table 3 shows results for the RMSE, RRSE, MAE, and R2 metrics for all evaluated techniques. It shows that support vector regression with linear kernel has the lowest RMSE, RRSE, MAE values followed by random forest regression, and second-degree polynomial support vector regression with the value of gamma = auto (1 / n_features). Instead there was a little difference with R^2, placing support vector regression first, followed by second-degree polynomial support vector regression with value of gamma = auto (i.e. 1 / n_features) and second-degree polynomial support vector regression with value of gamma = scale (i.e. 1 / (n_features * X.var())). Therefore, support vector regression with linear kernel had the best predictive ability on sample data. A similar result was reported by Kumar et al. [16]. The actual and predicted yield for sample validated models is shown in Fig. 3.

Table 4 shows the global validity of models. Global validity means extra sampler validity. Kumar at al. attempted to calculate 3, 4, and 5-fold cross validation scores for support vector machine for rice yield prediction [17]. In this study two-fold cross validation scores were calculated to test global validity. Results showed that support vector regression with a linear kernel has the highest cross validation score explaining 71% of variation with minimum mean standard error followed by second-degree polynomial support vector regression with a value of gamma = auto and random forest regression. KNN and support vector regression with radial basis function as a kernel function had negative crossvalidation scores. Fig. 4 presents the actual and predicted yield of cross validated models.

This work deals only with comparing the predictive ability of the above-mentioned machine learning models for Cluster bean crop only. Results may differ for other crops as every crop has different requirements.

4. CONCLUSION

Support vector regression with linear kernel was found to be the best-fitted model for predicting the yield as it had higher sample validity (98.31%) and global validity (71%). Production and the irrigated area had a higher positive impact and nitrogen consumption had the most negative impact on predicted yield. Data was a limitation of this study. Future research may be carried out with large and extended data sets.

ACKNOWLEDGEMENT

The author acknowledges the support from the Department of Agriculture Rajasthan, Jaipur, and particularly the Department of Agricultural Statistics, which provided datasets used in this study. We sincerely acknowledge the support of College of Agriculture, Swami Keshwanand Rajasthan Agricultural University, Bikaner for providing the research opportunity, institutional and administrative support.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Meftahizade Heidar, Hamidoghli, Yousef, Assareh, Mohammad, Javanmard, Majid. Effect of sowing date and irrigation regimes on yield components, protein and galactomannan content of guar (*Cyamopsis tetragonoloba* L.) in Iran climate. Australian Journal of Crop Science. 2017;11:1481-1487.
2. Gonzalez-Sanchez A, Juan FS, Waldo OB. Predictive ability of machine learning methods for massive crop yield prediction. Spanish Journal of Agricultural Research. 2014;12(02).
3. Roel A, Plant RE. Factors underlying yield variability in two California rice fields. Agronomy Journal. 2004;96(5):1481-1494.
4. Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE, et al. Random forests for global and regional crop yield predictions. PLoS One. 2016;11(6): e0156571. Available:https://doi.org/10.1371/journal.pone.0156571
5. Pavani S, Beulet AS. Heuristic prediction of crop yield using machine learning technique. International Journal of Engineering and Advanced Technologies. 2019;09(01):135 – 138.
6. Khosla E, Dharavath R, Priya R. Crop yield prediction using aggregated rainfall-based modular artificial neural network and support vector regression. Environ Dev Sustain. 2020;22:5687-5708
7. Drummond ST, Sudduth KA, Joshi A, Birrell SJ, Kitchen NR. Statistical and neural methods for site-specific yield prediction. T ASABE. 2003;46(1):5-14.
8. Devika B, Ananthi, B. Analysis of crop yield prediction using data mining technique to predict annual yield of major crops. International Research Journal of Engineering and Technology. 2018;05(12): 1460-1465.
9. Mishra S, Mishra D, Gour HS. Applications of machine learning techniques in agricultural crop production: A review paper. Indian Journal of Science and Technology. 2016;9(38):1-14.
10. Zhang L, Zhang J, Kyei-Boahen S, Zhang M. Simulation and prediction of soybean growth and development under field conditions. Am-Euras J Agr Environ Sci. 2010;7(4):374-385.
11. Smola A, Schölkopf B. A Tutorial on Support Vector Regression. Statistics and Computing. 2004;14(3):199-222.
12. Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. In: Advances in neural information processing systems (Mozer M, Jordan M, & Petsche T, eds), MIT Press, Cambridge, MA, USA. 1997:281-287.
13. Manjula E, Djodiltachoumy S. A model for prediction of crop yield. International Journal of Computational Intelligence and Informatics. 2017;6(4):298-305.
14. Noronha P, Divya J, Shruthi BS. Comparative study of data mining techniques in crop yield prediction. International Journal of Advanced Research in Computer and Communication Engineering. 2016;05(02): 132-135.
15. Porchilambi K, Sumitra P. Machine learning algorithms for crop yield prediction: A survey. Journal of Emerging Technologies and Innovative Research. 2019;06(03):112-116.
16. Kumar A, Kumar N, Vats V. Efficient crop yield prediction using machine learning algorithms. International Research Journal of Engineering and Technology. 2018; 05(06):3151-3159.
17. Kumar S, Kumar V, Sharma RK. Rice yield forecasting using support vector machine. International Journal of Recent Technology and Engineering. 2019;08(04):2588-2593.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/60327