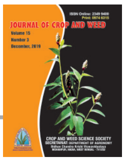
Journal of Crop and Weed, 18(1): 210-214 (2022)

ISSN- O : 2349 9400 ; P : 0974 6315



http://cwssbckv.org
www.cropandweed.com

Exploring the suitability of machine learning algorithms for crop yield forecasting using weather variables

M. VARMA, K. N. SINGH AND *A. LAMA

ICAR – Indian Agricultural Statistics Research Institute, New Delhi, 110012

Received : 12.01.2022 ; Revised: 22.03.2022 ; Accepted 28.03.2022

DOI: https://doi.org/10.22271/09746315.2022.v18.i1.1553

ABSTRACT

Crop yield forecast is valuable to many players in the agri-food chain, including agronomists, farmers, policymakers and merchants of commodities. Machine learning may be used to estimate crop yields, as well as to decide what crops to sow and what to do during the growing season. In present study Machine learning techniques such as Random Forest Regression and Support Vector Regression has been applied on three different datasets. Statistical indicators like Root Mean Square Error (RMSE), Mean Absolute Prediction Error (MAPE), and Mean Absolute Deviation (MAD) were used to compare the suggested models' forecasting performance. Also comparison has been done of both the machine learning techniques with the stepwise regression method. Support Vector regression was observed as the best machine learning technique. However, performance of the popular statistical approach (Stepwise regression) was found to be in between the two-machine learning algorithm.

Keywords**:** Machine learning, random forest, support vector regression, weather indices

Machine learning employs a data-driven or empirical modelling technique to detect meaningful patterns and relationships from input data and hence machine learning is a favorable choice for improving crop yield estimates. ML algorithms attempt to build a function that ties features or predictors to labels such as crop yield. Like statistical models, ML algorithms can use the findings of other approaches as features. Machine learning algorithms also have a variety of advantages, including the ability to simulate non-linear correlations between various data sources. ML could combine the strengths of earlier technologies, such as crop growth models and remote sensing, with data-driven modelling to provide realistic agricultural yield forecasts. Hence, machine learning (ML) techniques which has very few prior assumptions and are data driven provides great deal of flexibility for modelling and forecasting the crop yield.

ML is a practical technique for predicting agricultural yields based on different parameters. It is a branch of Artificial Intelligence (AI) that is dedicated on learning. By recognizing trends and relationships, machine learning (ML) can extract knowledge from datasets. The models should be trained with datasets that reflect previous experience-based outcomes. The parameters of the models are set during the training phase using past data, and the predictive model is constructed utilizing many features. The proposed techniques used have been tested to evaluate the performance using the testing dataset which we have been previously chosen. As, already mentioned, crop yield prediction is a complex phenomenon and has many underlining nonlinear patterns, such datasets are difficult to deal with stringent assumptions of the statistical models.

Random forests are an effective tool for forecasting. Due to Law of large numbers, Random forests do not over fit Breiman (2001). In compare to standard neural network, Support Vector Machine provides a number of advantages (Cristianini and Ricci, 2008). SVM (Support Vector Machine), and RF (Random Forest) are the better performing techniques for prediction of sugarcane yield (Bocca and Rodrigues, 2016). Combining machine learning with empirical domain knowledge increases predictive ability (Droesch, 2018). Small sample size is associated to higher classification accuracy (Vabalas *et al.,*2019). Also accurate machine learning models have been proposed with small datasets (Zhang and Ling, 2018).

Hence, machine learning techniques which has very few prior assumptions and are data driven, provides great deal of flexibility for modelling and forecasting the crop yield. Various researchers have applied different ML techniques for forecasting crop yield and have obtained satisfactory results (Chlingaryan *et al*., 2018; Droesch, 2018 and Gopal and Bhargavi, 2019). Many of the ML techniques are already developed and are used like Random Forest, Support vector machine, KNN, Logistic Regression, K-Means *etc.*On the basis of temperature, rainfall, season, and area various machine learning algorithms for predicting crop yield have been presented (Nigam *et al.,*2019).

MATERIALS AND METHODS

Machine learning algorithm started with the collection of data. After collection data preparation was done under which the dataset was divided under 2 parts out of which bigger part was used for training purpose

Email: chllm6@gmail.com

and the smaller part was used for validation purpose. With the help of training dataset, the model was trained and after training the validation of the predictive model has been done with help of validation dataset. After validation the model will be ready for prediction.

Data Gathering: Collection of weekly weather data containing different weather parameter such as minimum temperature, maximum temperature, relative humidity, total precipitation, mean temperature, and atmospheric pressure has been done. Also collection of the crop yield data has been done for the corresponding districts.

Data preparation: For the Medak district of Telangana state rice yield dataset, 30 weather indices have been formed. For the Baran district of Rajasthan state wheat yield dataset also 30 indices have been formed and for the Jalandhar district of Punjab state wheat yield dataset, 56 weather indices have been formed. These weather indices have been used for further analysis. By applying the function for the formation of weather indices, different number of weather indices would be obtained for different datasets.

Weather indices would be formed by using expression-

$$Y=A_0+\sum_{i=1}^{p}\sum_{j=0}^{1}a_{ij}Z_{ij}+\sum_{i'\,i=1}^{p}\sum_{j=0}^{1}a_{ii'}Z_{ii'}+C+\varepsilon$$

Where,

$$Z_{ij}=\sum_{w=1}^{m}r_{iw}^{j}X_{iw}$$

$$Z_{ii'j}=\sum_{w=1}^{m}r_{ii'w}^{j}X_{iw}X_{i'w}$$

Where, $r_{iw}/r_{ii'j}$ is correlation coefficient of the yield with ith weather variable / product of ith and i'th variables in wth week, m is week of forecast p is the number of weather variables used (Agrawal and Mehta, 2007).

Number of indices formed from n weather variables

$$k=2\left[n+\binom{n}{2}\right]$$

So, for n = 2, no of indices will be 6
n = 3, no of indices will be 12
n = 5, no of indices will be 30
(Singh *et al.,* 2019)

We also had to divide the data into two sections. The majority of the dataset would be used to train our model in the first section. The second section would be used to assess the performance of our trained model.

Choosing a model: The selection of a model is the next step in our process. Over the years, researchers and data scientists have developed a variety of models. Some are better suited to image data, while others are best suited to sequences, numerical data, or text-based data. In this study we were using two models Random Forest Regression model and Support Vector Regression model.

Training: In this step, we used our data to gradually improve our model's predictive ability. Through training, the algorithms were used to gain the experience which was formed by observations of training dataset.

Evaluation: The test set is a collection of data used to evaluate the model's performance using a performance metric.

Parameter Tuning: Once evaluation was completed, it is possible that we wished to see whether there was any way to improve training. This could be done by tuning our parameters. While training the forecasting models under study, a few parameters have been assumed which were verified and different values have been tried.

Prediction: Data is used in machine learning to answer the questions. So, inference or prediction is the stage when we get to answer certain questions. This is the conclusion of all of our efforts, and so, the value of machine learning was realized at this stage.

Data description

In the present study, work has been done on 3 datasets, out of which first data has been collected for the rice crop yield and different weather parameters (Maximum and minimum temperature, rainfall, Relative humidity (I&II)) for Medak district of Telangana provided by IMD, New Delhi. For dataset 2 wheat crop yield with different weather parameters has been collected for Baran district of Rajasthan. Dataset 3 also contains the wheat crop yield data with different weather parameters which has been taken for the Jalandhar district of Punjab. Weather data was taken from the website - https://rds.ncmrwf.gov.in/ .

Random Forest Regression

Random forest is a flexible, easy-to-use machine learning method that, in most cases, delivers good results even without hyper-parameter tuning. Because of its simplicity and diversity, it is also one of the most often used algorithms. Random forest is supervised machine learning algorithm. It builds a "forest" out of an ensemble of decision trees, which are generally trained using the "bagging" process. The bagging method's general concept is that combining several learning models improves the final outcome (Huang, 2014).

Procedure

- Pick k data points at random from the training dataset.

- Create a decision tree based on these k points.
- Select the number N of trees wish to be constructed and repeat the procedures above.
- Make each of your N-tree predict the value of y for a new data point, then assign the new data point to the average of all predicted y values.

Form of the regression trees model –

$$f(X)=\sum_{m=1}^{M}\mathrm{c}_{\mathrm{m}}\cdot 1_{(\mathrm{X}\in\mathrm{R}_{\mathrm{m}})}$$

Where, $R_1, R_2, \ldots, R_M$ represent a partition of feature space.

Support Vector Regression

To predict discrete values, the supervised learning algorithm Support Vector Regression is employed. The Support Vector Regression and the Support Vector Machine both function on the same principle. The core objective of SVR is to find the best-fitting line. In SVR, the hyperplane with the greatest number of points is the best fit line. Unlike other regression models, the SVR seeks to fit the best line within a threshold value rather than minimising the difference between the real and predicted values. The threshold value is the distance between the hyperplane and the boundary βine.

Fitting of Support Vector regression $f(X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ can be expressed as-

$$minimize\left\{\sum_{i=1}^{n} max[0,1-y_i f(x_i)]+\underline{l}\sum_{j=1}^{p}{}_{j}^{2}\right\}$$

(Vapnik, 1998)

RESULTS AND DISCUSSION

Weather indices formed by the method shown in the previous section have been directly utilized for applying different machine learning algorithms. For all the three datasets, Random Forest Regression and Support Vector Regression techniques have been employed for the prediction of crop yield. As the stepwise regression was found to be well developed and efficient method in literature, a comparison of RF and SVR have been also done with stepwise regression.

In present study 500 tress have been used for all the three datasets. The fitting of the regression model was done for all the three datasets as mentioned in Fig. 1, 2 and 3 for Medak, Baran and Jalandhar district respectively.

For Support Vector Regression linear kernel function was used for all three datasets. The cost of constraint violation was taken as 1. For tuning SVR model the epsilon values have been taken as 0,0.1, 1 and the cost function has been taken in the range 1 to 100. The fitting of SVR ws done with all three datasets as mentioned in

Table 1: Comparison of ML techniques with in-sample data of Medak district of Telangana rice yield (Kg ha^{-1}) data

Algorithm	MAD	RMSE	MAPE
Random Forest	83.70	93.94	3.46
SVR	101.86	71.65	4.45

SVR – Support vector regression, RMSE - Root Mean Square Error, MAPE - Mean Absolute Prediction Error, MAD- Mean Absolute Deviation

Table 2: Comparison of ML techniques with in-sample data of Baran district of Rajasthan wheat yield (Kg ha^{-1}) data

Algorithm	MAD	RMSE	MAPE
Random Forest	531.16	440.74	22.31
SVR	146.01	105.67	5.90

Table 3: Comparison of ML techniques with in-sample data of Jalandhar district of Punjab wheat yield (Kg ha^{-1}) data

Algorithm	MAD	RMSE	MAPE
Random Forest	713.68	480.31	22.27
SVR	608.65	536.40	17.31

Table 4: Comparison of predicted values by ML techniques with actual values for Medak district of Telangana rice yield (Kg ha^{-1}) dataset.

Years	Actual values (Kg ha^{-1})	Predicted values (Kg ha^{-1})	
		Random forest	SVR
2013	3627	3307.80	2831.26
2014	3168	3175.35	2832.48
2015	2973	2835.03	2636.66
2016	3063	2847.25	2386.76
2017	3924	3315.11	3083.37

SVR – Support Vector Regression

Table 5: Comparison of predicted values by ML techniques with actual values for Baran district of Rajasthan wheat yield (Kg ha^{-1}) dataset

Years	Actual values (Kg ha^{-1})	Predicted values (Kg ha^{-1})	
		Random forest	SVR
2013	3600	3258.77	3458.55
2014	3386	2987.52	3287.42
2015	3831	3335.25	3654.68
2016	4207	3854.85	3855.05

SVR – Support vector regression

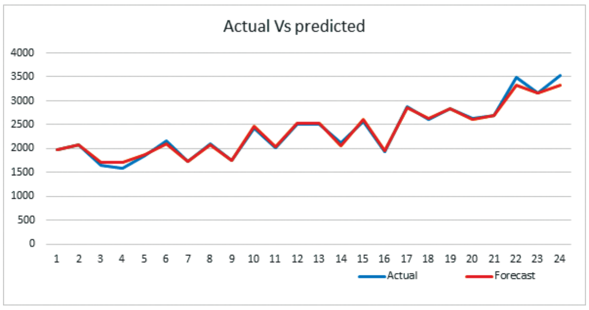


Fig. 1: Fitting of Random forest for Medak district of Telangana rice yield (Kg ha^{-1}) data

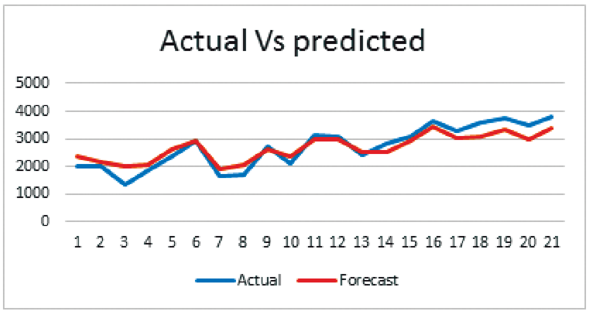


Fig. 2: Fitting of Random forest for Baran district of Rajasthan wheat yield (Kg ha^{-1})

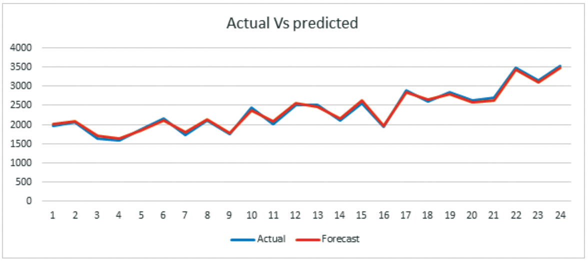


Fig. 3: Random Forest fitting for Jalandhar district of Punjab wheat yield (Kg ha^{-1}) data

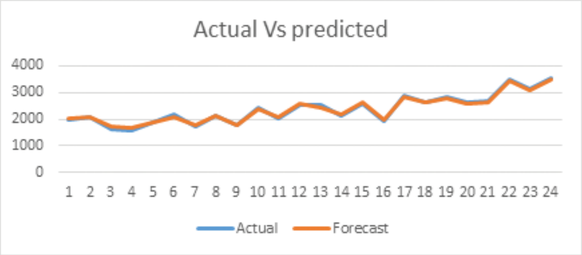


Fig. 4: Fitting of SVR for Medak district of Telangana rice yield (Kg ha^{-1}) data

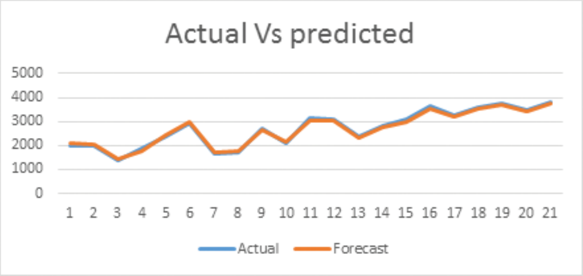


Fig. 5: Fitting of SVR for Baran district of Rajasthan wheat yield (Kg ha^{-1}) data

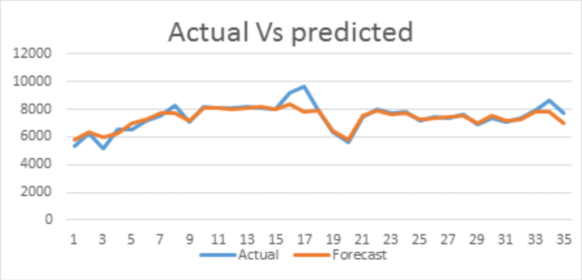


Fig. 6: Fitting of SVR for Jalandhar district of Punjab wheat yield (Kg ha^{-1}) data

fig. 4, 5 and 6 for Medak, Baran and Jalandhar district, respectively.

Comparison on model fitting on the basis of in-sample.

For the Random Forest Regression and Support Vector Regression, the forecasting accuracy of the models with in-sample dataset have been compared which is mentioned in the Table 1, 2, and 3 for the Medak, Baran and Jalandhar district, respectively.

Comparison between predicted values

The tables containing the actual values of the testing dataset and their predicted values obtained through both ML techniques have been mentioned below for all three datasets in Table 4, 5 and 6.

Comparison of the prediction accuracy

The comparison between the prediction accuracy of RF, SVR and Stepwise regression has been carried out by MAD, RMSE, and MAPE measures mentioned in Table 7,8 and 9. It has been observed that, for Medak district of Telangana rice yield data all the three measures have been observed minimum in the case of Random Forest Regression. Whereas, for Baran district of Rajasthan wheat dataset and for Jalandhar district of Punjab wheat yield data all the three measures have been observed minimum in the case of SVR.

Conclusion

Present study has been carried out on 03 different datasets of varying lengths (25, 29 and 40). 30 weather indices were obtained each for Medak district Telangana rice yield dataset and Baran district Rajasthan wheat yield datasets and 56 for Jalandhar district Punjab wheat dataset.When machine learning techniques were applied on datasets, prediction accuracy of the SVR (Support Vector Regression) was to be found higher as compared to the Random Forest technique, except for Medak district Telangana rice yield. However, the popular

Table 6: Comparison of predicted values by ML techniques with actual values for Jalandhar district of Punjab wheat yield (Kg ha^{-1}) dataset

Years	Actual values (Kg ha^{-1})	Predicted values (Kg ha^{-1})	
		Random forest	SVR
2014	4469	3695.53	3881.75
2015	3356	3232.43	3790.27
2016	4654	4016.02	4167.05
2017	4606	3358.31	3848.64
2018	4733	4011.75	4394.44

**SVR – Support vector regression*

Table 7: Comparison of prediction accuracy of ML techniques for Medak district of Telangana rice yield (Kg/ha) data

Algorithm	MAD	RMSE	MAPE
Random Forest	257.83	328.11	7.25
SVR	596.90	636.06	17.47
Stepwise Regression	405.82	460.10	12.77

**SVR – Support vector regression*

Table 8: Comparison of prediction accuracy of ML techniques for Baran district of Rajasthan wheat yield (Kg ha^{-1}) data

Algorithm	MAD	RMSE	MAPE
Random Forest	396.90	401.560	10.63
SVR	192.07	214.87	4.95
Stepwise Regression	571.14	654.49	15.50

*SVR – Support vector regression

Table 9: Comparison of prediction accuracy of ML techniques for Jalandhar district of Punjab wheat yield (Kg ha^{-1}) data.

Algorithm	MAD	RMSE	MAPE
Random Forest	700.79	787.08	15.40
SVR	520.88	540.15	12.03
Stepwise Regression	744.670	961.717	9.73

*SVR – Support vector regression

statistical approach (Stepwise regression) performance was found to be in between the two-machine learning algorithm.

ACKNOWLEDGEMENT

We want to express our respect and gratitude to Indian Meteorological Department for providing the dataset and to the Director ICAR – IASRI for providing the facilities in the institute. First author would like to convey deep sense of gratitude and appreciation to whole advisory committee for their continuous guidance and support and also thankful to ICAR for providing the ICAR – JRF fellowship.

REFERENCES

Agrawal, R. and Mehta, S. 2007. Weather based forecasting of crop yields, pests and diseases -IASRI Models. *J. Indian Soc. Agric. Stat.*, **61**: 255-263.

Bocca, F.F. and Rodrigues, L.H.A. 2016. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Computers and Electronics in Agriculture*, **128**: 67-76.

Breiman, L. 2001. Random forests. *Machine learning*, **45**: 5-32.

Cristianini, N. and Ricci, E. 2008. Support Vector Machines. *Encyclopedia of Algorithms. Springer,* 928-932

Chlingaryan, A., Sukkarieh, S. and Whelan, B. 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, **151**: 61-69.

Droesch, A.C. 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ.Res. Letters*, **13**: 1-12.

Gopal, P.M. and Bhargavi, R. 2019. Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. *Applied Artificial Intelligence,***33**: 621-642.

Huang, J. Z. 2014. *An Introduction to Statistical Learning: With Applications in R.* Springer

Nigam, A., Garg, S., Agrawal, A. and Agrawal, P. 2019. Crop yield prediction using machine learning algorithms. In *2019 Fifth International Conference on Image Information Processing (ICIIP)* (pp. 125-130). IEEE.

Singh, K.N., Singh, K.K., Kumar, S., Panwar, S. and Gurung, B. 2019. Forecasting crop yield through weather indices through LASSO. *Ind. J. Agric. Sci.*, **89**: 540-544.

Vabalas, A., Gowen, E., Poliakoff, E. and Casson, A. J. 2019. Machine learning algorithm validation with a limited sample size. *PloS One*, **14**(11), e0224365.

Vapnik, V.N. 1998. *Statistical Learning Theory* (1st ed.), Wiley-Interscience.

Zhang, Y. and Ling, C. 2018. A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*, **4**, 1-8.