Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

# Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms

**Maya Gopal P. S. & Bhargavi R.**

**To cite this article:** Maya Gopal P. S. & Bhargavi R. (2019) Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms, Applied Artificial Intelligence, 33:7, 621-642, DOI: 10.1080/08839514.2019.1592343

**To link to this article:** https://doi.org/10.1080/08839514.2019.1592343

APPLIED ARTIFICIAL INTELLIGENCE
2019, VOL. 33, NO. 7, 621–642
https://doi.org/10.1080/08839514.2019.1592343

Taylor & Francis
Taylor & Francis Group



# Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms

Maya Gopal P. S. and Bhargavi R.

School of Computing Science and Engineering, VIT University, Chennai, India

**ABSTRACT**

The rapid innovations and liberalized market economy in agriculture demand accuracy in Crop Yield Prediction (CYP). In accurate prediction, machine learning (ML) algorithms and the selected features play a major role. The performance of any ML algorithm may improve with the utilization of a distinct set of features in the same training dataset. This research work evaluates the most needed features for accurate CYP. The ML algorithms, namely, Artificial Neural Network, Support Vector Regression, K-Nearest Neighbour and Random Forest (RF) are proposed for better accuracy. Agricultural dataset consists of 745 instances; 70% of data are randomly selected and are used to train the model and 30% are used for testing the model to assess the predictive ability. The results show that the RF algorithm reaches the highest accuracy by means of its error analysis values for all the distinct feature subsets using the same training agricultural data.

## Introduction

Crop production is influenced by various parameters including climate, soil quality and fertilizer (Henryson et al. 2018; Taylor et al. 2018; Ximena et al. 2017). Crop Yield Prediction (CYP) mainly depends on two major feature sets. One set of data contain the land usage, land preparation, applied fertilizers, and the methods of irrigation, which depends on the farmer. The other set of data contain environmental features such as temperature, rainfall, and solar radiation which are controlled by nature (Shine et al. 2018). However, collecting the data which are relevant to the CYP is very time-consuming and tedious process. Feature selection algorithms provide the suitable features which are most related to the CYP based on their own selection criteria. Applying the ML algorithm and tuning their parameters based on the feature set make an accurate prediction. Researchers are working toward developing efficient methods to evaluate the prediction accuracy based on the data which they collected (Chlingaryan, Sukkarieh, and Whelan 2018; Gonzalez-Sanchez, Frausto-Solis, and Ojeda-Bustamante 2014; Johnson et al. 2016). As a result, the data-driven models have gained popularity and

**CONTACT** Maya Gopal P. S. mayagopal.ps2016@vitstudent.ac.in School of Computing Science and Engineering, VIT University, Chennai, India

found applications for CYP using classical statistical and machine learning (ML) methods. ML approaches such as Artificial Neural Network (ANN), Support Vector Regression (SVR), K-Nearest Neighbour (KNN), and Random Forest (RF) which are parametric or nonparametric in nature and are heavily dominating the CYP in different agricultural data sets.

Multilayer perceptron neural network of feed-forward model has been used to predict biomass yield using input features such as climate, soil texture, crop rotation, soil structure, crop residues management method, depth of tillage, tillage method, the amount of (nitrogen, phosphorus and potash) fertilizers consumed, and efficiency of water usage. The model has a determination coefficient $R$ of 90% for tillage method (Mehnatkesh et al. 2012; Mobarake et al. 2014). Tracing the relationship between the yield of rain-fed corn and the features such as soil, climate, and management using a feed-forward, back propagation ANN model showed that the key factor was the amount of rainfall recorded during the crop seasons (Liu, Goering, and Tian 2001). Estimating rain-fed wheat yield using the independent features like elevation, slope, slope direction, curvature, specific contributing area, and moisture index, made a comparison of Spatial Analysis Neural Network (SANN) method with the Multiple Linear Regressions (MLRs) technique and shows that, by using features relating to topography as the inputs the Root Mean Square Error (RMSE) in SANN with five independent variables was 0.59 and MLR with four or five independent variables was 0.72 (Green, Salas, and Ana Martinez 2007). Ji et al. predicted rice crop yield using artificial neural networks in China's province of Fujian, also comparing the neural network models with MLR model, using data on climatic variables and the local rainfall, shows that the neural network model predicted better performance when compared with the regression model. $R^2$ and RMSE obtained for the ANN model were 0.67 and 891, respectively, as against 0.52 and 1977 for the regression model (Ji et al. 2007). Cartesian Genetic Programming technique Polak-Ribiére is used to analyze six artificial neural network models by different training techniques to select a model with Minimum Associated Relative Error (MARE) for predicting the wheat production in Iran (Ghodsi et al. 2012). The inputs were the amount of rainfall, guaranteed purchase price, the area under cultivation, subsidies, the insurance rates, imports, population, added value of agricultural sector, and the output was wheat production. In a bid to spot the most accurate technique, comparisons were carried out among regression models for CYP. Drummond et al. (2003) have compared the classical statistical model with ANNs (Drummond et al. 2003). Similar work was carried out by Fortin et al. (2011). Ruß made a comparison of ANNs, regression trees and support vector regression (Ruß 2009). Even as it is highly site-dependent, neural networks have been widely popularised as a robust model, fetching valid results for reliable CYP. On the other hand, support vector regression models have displayed a higher degree of accuracy when compared to the ANN and regression trees for selected crop datasets. The researchers worked with the features like rainfall, temperature,

fertilizers, total area cultivated and method of irrigation to analyze the CYP (Brown et al. 2018; Gonzalez-Sanchez, Frausto-Solis, and Ojeda-Bustamante 2014; Satir and Berberoglu 2016). One of the recent articles, Gonzalez-Sanchez et al. (2014), presented a comparative study of MLR and ANN, SVR, M5-Prime, KNN ML techniques for CYP using 10 crop datasets. To validate the models they used four accuracy metrics and the results showed that M5-Prime achieved the lowest errors across the produced crop yield models (Gonzalez-Sanchez, Frausto-Solis, and Ojeda-Bustamante 2014). Nari and Yang-Won (2016) applied SVM, RF, extremely randomized trees and Deep Learning (DL) to estimate the corn yield in Iowa State and showed that DL provided more stable results by overcoming the overfitting problem (Nari and Yang-Won 2016).

In addition, the best of our understanding, very few works were published to evaluate and compare the ANN, SVR, KNN, and RF with distinct feature sets for the same training data set. The two main objectives of the current work are: (a) to evaluate the prediction accuracy of the four ML algorithms, ANN, SVR, KNN and RF and (b) to assess the importance of distinct feature sets on ML algorithms. The feature selection algorithms used to identify the distinct feature sets are Forward Feature Selection (FFS) algorithm, Correlation-based Feature Selection (CBFS) algorithm, Variance Inflation Factor (VIF) algorithm, and Random Forest Variable Importance (RFVarImp) algorithm. The statistical programming language R is used to analyze the data due to its increased sustainability.

## Data Set

The historical data of the paddy crop in the State of Tamil Nadu, in the southern part of India, which is located in the tropical region, are used. In this research work, the data were collected from the meteorological department of India, agricultural department of Tamilnadu and the statistical department of Tamilnadu. The features considered for this research work are listed in Table 1. The used features are planting area (ha), number of tanks, number of tube wells and open wells used for irrigation and canal length in meters for irrigation, amount of fertilizers such as nitrogen, phosphorus and potash (kg) consumed, seed quantity for the planting area (kg), cumulative rainfall (mm), cumulative global solar radiation (kWh $m^{-2}$), maximum, average and minimum temperatures (°C). The collected data are cleaned and rescaled with range between 0 and 1 in order to find the accurate prediction.

## Feature Selection Algorithm

The feature selection algorithms help, feeding in only those features that are relevant in the predictive algorithms (Oreski, Oreskib, and Klicek 2017). The distinct feature subsets selected from the feature selection algorithm are used for CYP. Instead of a complete set of features, feature subsets give better results for the

**Table 1.** Description of the dataset.

| Feature ID | Feature type | Description |
|---|---|---|
| CL | Predictor | Canal length used for irrigation in meter |
| TK | Predictor | Total number of tanks used for irrigation |
| TW | Predictor | Total number of tube wells used for irrigation |
| OW | Predictor | Total number of open wells used for irrigation |
| AH | Predictor | Total land area used for cultivation in hectare |
| NF | Predictor | Total amount of nitrogen used for cultivation for the year |
| PF | Predictor | Total amount of phosphate used for cultivation for the year |
| KF | Predictor | Total amount of potash used for cultivation for the year |
| SD | Predictor | Total quantity of seed used for cultivation in kg |
| RainF | Predictor | Average rainfall for the year in mm |
| AT | Predictor | Average daily mean temperature registered for the year |
| TMin | Predictor | Average of daily minimal temperature registered for the year |
| Tmax | Predictor | Average of daily maximum temperature registered for the year |
| SR | Predictor | Average of accumulated daily radiation in the year |
| PD | Target/response | Total production of the year in ton |

same algorithm with less computational time. The main reason for using feature selection is that (i) it enables the ML algorithm to train faster, (ii) reducing the complexity of a model and (iii) making it easier for interpretation. It also improves the accuracy of a model if the right subset is chosen and reduces overfitting. Feature selection has the potential to play an important role in the agriculture domain, with the production depending on land use, irrigation applied, fertilizer applied and weather parameters. Feature selection algorithm is applied to spot the key features which have a strong correlation with crop yield (Bijanzadeh et al., 2010). The selected features of different feature selection algorithms are listed in Table 2.

## *FFS Algorithm*

FFS involves iteration beginning with a null feature set, including the best feature in each step, satisfying the certain criterion, in pursuit of the desired features. The FFS algorithm is based on the Akaike Information Criterion (AIC) value for feature selection. This algorithm selects the features on each step along with previous features which are selected since these are considered with a new one. The forward method initiates with a null set and expands it. The features selected

**Table 2.** Features selected by each feature selection method.

| Features / feature selection methods | AH | CL | TK | TW | OW | SD | RainF | AT | TMin | Tmax | SR | NF | PF | KF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Forward feature selection | √ | √ | √ | | √ | | | | | √ | | | | |
| Correlation based feature selection method | √ | √ | √ | √ | √ | | | √ | | √ | √ | | | |
| Random Forest Var. Imp | √ | | √ | | √ | | | | | | √ | √ | √ | √ |
| VIF | √ | √ | √ | √ | √ | | √ | √ | √ | √ | √ | | | |

**Table 3.** Forward feature selection procedure by using AIC.

| Feature subset | AIC value | Selection procedure |
|---|---|---|
| { } | −1569.58 | |
| {AH} | −2462.11 | ↓ |
| (AH, OW} | −2516.1 | ↓ |
| {AH, OW, TK} | −2528.66 | ↓ |
| {AH, OW, TK, CL} | −2533.53 | ↓ |
| {AH, OW, TK, CL,Tmax} | −2534.84 | Stop |
| {AH, OW, TK, CL, Tmax, NF} | −2534.8 | ↑ |

by the algorithm satisfies the lowest AIC value for forward feature selection. While adding one more feature on this set, the AIC value is increased. So at this point, the selection procedure is stopped. The selection procedure and the feature selected are listed in Table 3. It selects the features area in a hectare, number of open wells, number of tanks, canal length, and maximum temperature used for the crop.

### CBFS Algorithm

CBFS involves a heuristic evaluation function based on correlation, ranking feature subsets towardwhich the evaluation function has a bias. The CBFS is calculated by using $merit_s = \frac{K\overline{r_{cf}}}{\sqrt{K+K(K-1)\overline{r_{ff}}}}$, where $K$ is the number of features in the subset, $\overline{r_{cf}}$ is the average correlation between each feature in $S$ and output variable $C$, $\overline{r_{ff}}$ is the average feature to feature pairwise correlation between the features in $S$. This algorithm gives the features which are highly correlated with production. The CBFS algorithm generates the correlation matrix. For the dependent feature PD, all the possible independent feature subsets are generated and their score is calculated. Based on the highest score the subsets were selected. The highest score subset contains the features area in hectare, number of open wells, number of tanks, maximum temperature, fertilizer such as nitrogen, phosphorus, and potash (kg) consumed, and seed quantity. The correlation matrix for the available data is shown in Figure 1.

### VIF Algorithm

VIF method, which is used to remove the correlated independent features, is extremely quick, using a one-pass search over the predictors. It is a computationally efficient method of testing each potential predictor for addition to the model. VIF regression probably avoids model over-fitting. VIF is calculated by using the formula $VIF = \frac{1}{1-R_i^2}$. VIF-based feature selection algorithm checks the collinearity among the independent features and selects the independent features with less colinearity between them. The algorithm checks the colinearity of the features individually and it requires more computational time. Among the independent features in the available data AH, SD, NF, PF and KF are colinear. So

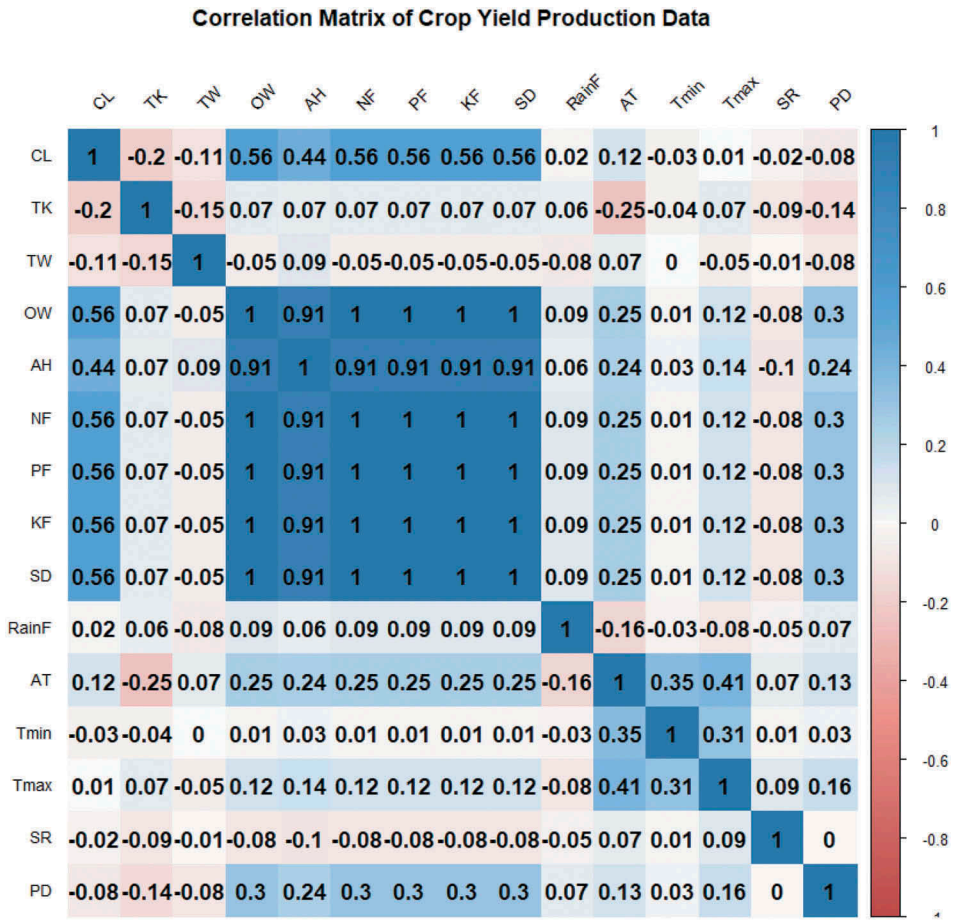


**Figure 1.** Correlation matrix of crop yield production data.

only one independent feature AH is selected from the feature set and all other features were removed from the feature set. The other features such as CL, TK, TW, OW, RainF, AT, Tmin, Tmax, and SR are selected since the non-colinearity of the features. The features which are selected by the VIF algorithm are planting area (ha), number of tanks, number of tube wells, and open wells used for irrigation and canal length in meters for irrigation, cumulative rainfall (mm), cumulative global solar radiation (kWh $m^{-2}$), maximum, average, and minimum temperatures (°C). The features and its VIF value are given in Table 4.

## *RFVarImp Algorithm*

Random forests is an ensemble learning method for classification and regression. These operate by constructing a multitude of decision trees at training time and delivering the class, i.e., the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests use a modified tree learning

**Table 4.** Features and its VIF value.

| Feature | VIF value |
|---|---|
| CL | 1.405580 |
| TK | 1.998218 |
| TW | 1.463954 |
| OW | 1.093255 |
| RainF | 1.055797 |
| AT | 1.544300 |
| Tmin | 1.147031 |
| Tmax | 1.269596 |
| SR | 1.030926 |
| AH | 2.285810 |

**Table 5.** Features and its node purity based on RFVarImp.

| Feature | IncNodePurity |
|---|---|
| CL | 0.4251797 |
| TK | 0.8106956 |
| TW | 0.3285912 |
| OW | 0.7854444 |
| AH | 4.0762534 |
| RainF | 0.3095985 |
| AT | 0.4484184 |
| Tmin | 0.1812348 |
| Tmax | 0.1778059 |
| SR | 0.1823419 |
| NF | 4.3358796 |
| PF | 4.5037572 |
| KF | 4.6039728 |
| SD | 4.0677608 |

algorithm that selects, a random subset of the features, at each candidate split in the learning process. This algorithm selects the variables based on their importance with respect to the response variable production and the features are selected based on the node purity. In this work, the node purity threshold is set as a median of the node purity values. The node purity value above the median value feature is selected. The features selected by RFVarImp are area in hectare, number of open wells, amount of fertilizers such as nitrogen, phosphorus, and potash (kg) consumed, seed quantity for the planting area (kg). The features and its node purity are listed in Table 5.

## ML Algorithms

ML algorithms build the computers to automatically improve the efficiency. The ML techniques focus on the predictive accuracy of models rather depending on the data modeling in statistics (Breiman 2001). Without or minimal human intervention, the ML gives better decision-making support. In this research work,

the ML algorithms, namely, ANN, SVR, KNN, and RF are used to analyze the CYP. The aforementioned algorithms, the tuning parameters play a major role in producing high prediction accuracy. Each algorithm has its own tuning parameters and its procedure for tuning. Based on different crop feature subsets, the tuning parameter values of ML algorithms may vary to achieve optimum accurate prediction. The predicted results under the optimal parameter of each algorithm are used for comparison. The algorithm of predictive algorithm with distinct feature subsets is given below.

**Algorithm**

Predictive algorithm with distinct crop feature subset

Input : Number of features X : = $\{x_1,x_2,x_3,\ldots\ldots\ldots\ldots x_n\}$
Output : Predicted yield value Y

Step 1 : Input the numer of features which are correlated with crop yield .
X : = $\{x_1,x_2,x_3,\ldots\ldots\ldots\ldots\ldots,x_n\}$

Step 2 : Apply the feature selection algorithm.(FSA)

Step 3: If FSA: = Forward FSA then goto step 6
Elseif FSA: = CBFS(Correlation Based Feature Selection) then goto step 7
Elseif FSA : = VIF(Varience Inflation Factor) then go to step 8
Else FSA : = RFVarImp(Random Forest) then go to step9

Step 4: Apply predictive algorithm and then find the predicted yield

Step 5: If predictive algorithm : = ANN or SVR or RF or KNN then
/* ANN – Artificial Neural Network, SVR -Support Vector Regression
RF – Random forest, KNN – k-Nearest Neighbour */

{
Apply distinct feature subsets(Xsubset)
Compute predicted value Y
Calculate the accuracy using performance metrices

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n}}$$

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

$$MAE = \left(\frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{(n)(\bar{y})}\right)$$

}

Step 5: End

Step 6: Procedure FFSA

$X : = \{x_1,x_2,x_3,\ldots x_n\}$

Calculate AIC value for each features

$$AIC := AIC = N\ln(\frac{SS_{error}}{N}) + 2K$$

If xi : = lowest AIC value then x subset = {xi}

Repeat

{

i : = 1

Calculate AIC value for each subset of X with i + 1 features when i features are are already exists in the subset which has lowest AIC value.

i = i + 1

Until Xsubset AIC value is greater than previous subset

Return $X_{subset} : = \{x_1,x_2,x_3,\ldots x_n\}$

Step 7: Procedure CBFS

$X : = \{x_1,x_2,x_3,\ldots x_n\}$

Generate the nxn correlation matrix for the entire features space.

Calculate the score of each feature subset by using

$$merit_s = \frac{K\overline{r_{cf}}}{\sqrt{K + K(K-1)\overline{r_{ff}}}},$$

Select the feature subset $X_{subset}$ which is having the highest score

Return $X_{subset} : = \{x_1,x_2,x_3,\ldots x_n\}$

Step 8: Procedure VIF

$X : = \{x_1,x_2,x_3,\ldots x_n\}$

Calculate VIF value for each feature $VIF = \frac{1}{1-R_i^2}$

Check the colinearity between the features in the set based on its VIF value.

If VIF > median of VIFs the remove the feature
Else select the feature.

Return $X_{subset}$ : = $\{x_1,x_2,x_3,\ldots x_n\}$

Step 9: Procedure RFVarImp
$X = \{x_1,x_2,x_3,\ldots x_n\}$

Calculate the importance of each feature by using random forest algorithm .
Find the node purity value
If node purity > median of node purity value the select the feature.

Return $X_{subset}$ : = $\{x_1,x_2,x_3,\ldots x_n\}$

### *Artificial Neural Network*

ANN is a computational model which mimic the human nervous system. The ANN is commonly applied to predict crop yield (Akbar et al. 2018; Monisha Kaul and Hill 2005; Torkashvand, Ahmadi, and Nikravesh 2017). It has three layers, namely, input layers, hidden layers, and output layer. The input layer consists a number of neurons which are equivalent to the number of input features, and the output layer has only one neuron which is crop yield. In this work, feed-forward neural network with backpropagation training algorithm is applied to find accurate crop yield. The number of input neurons differs based on feature sets which are obtained by using feature selection algorithms. The only tuning parameter of the algorithm is the number of hidden neurons in order to achieve better prediction. The number of hidden neurons may vary based on the number of input features.

Based on the number of features the number of hidden layer neuron is changed in order to obtain the accurate prediction. In this work, we applied a trial-and-error method, to select the number of hidden neurons. (Hill 2010; Shamseldin 1997). In RFVarImp, the number of features selected is 7. Therefore, four neurons are selected for the hidden layer. The neural network for RFVarImp-based feature subset is shown in Figure 2. In CBFS, the number of features selected is 10. Therefore, seven neurons are selected for the hidden layer. The neural network for CBFS-based feature subset is shown in Figure 3. In VIF, the number of features selected is 10. Therefore, seven neurons are selected for the hidden layer. The neural network for VIF-based feature subset is shown in Figure 4. In FFS, the number of features selected is 5. Therefore, three neurons are selected for the hidden layer. The neural network for FFS-based feature subset is shown in Figure 5.

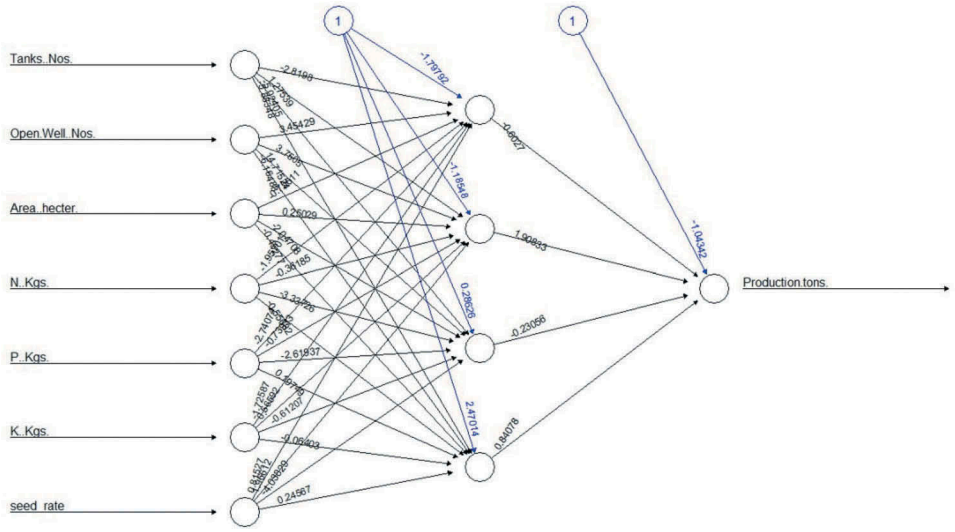


**Figure 2.** Neural network for RFVarImp-based feature subset.

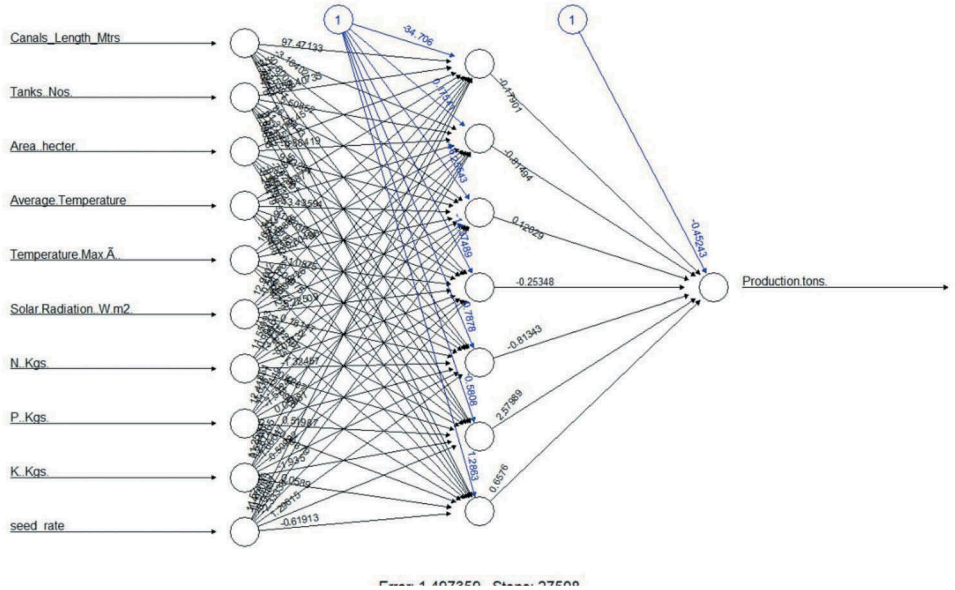


**Figure 3.** Neural network for CBFS-based feature subset.

## *Support Vector Regression*

SVR is commonly used in crop yield prediction (Gu et al. 2016; Ying-Xue, Huan, and Li-Jiao 2017). One of the advantages of this method is that mathematical analysis is relatively easier because nonlinear problems related to the input space are expressed by being matched with linear problems of high-dimension feature space (Hearst et al. 1998). In SVR, radial basis function kernel is commonly set to

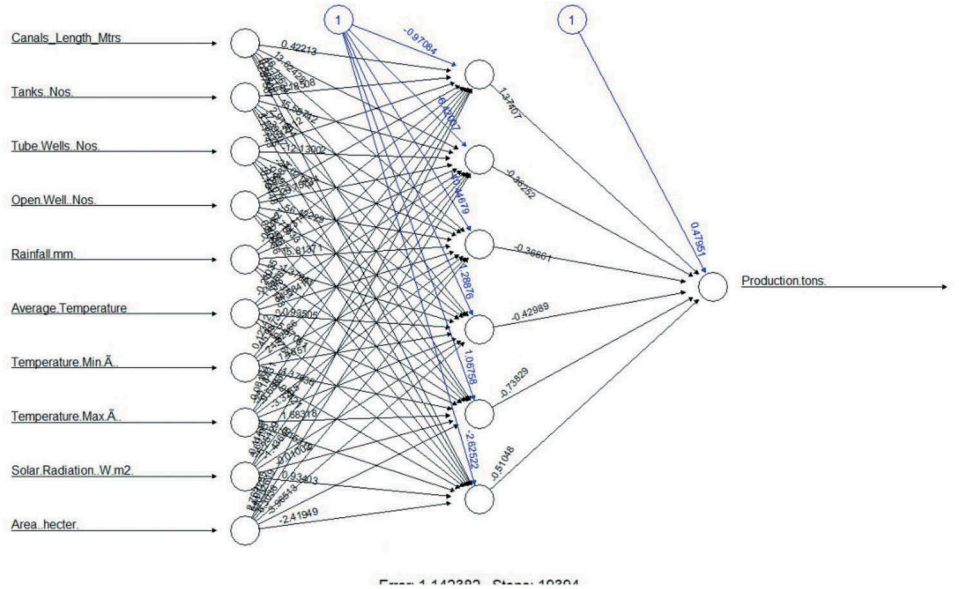


**Figure 4.** Neural network for VIF-based feature subset.

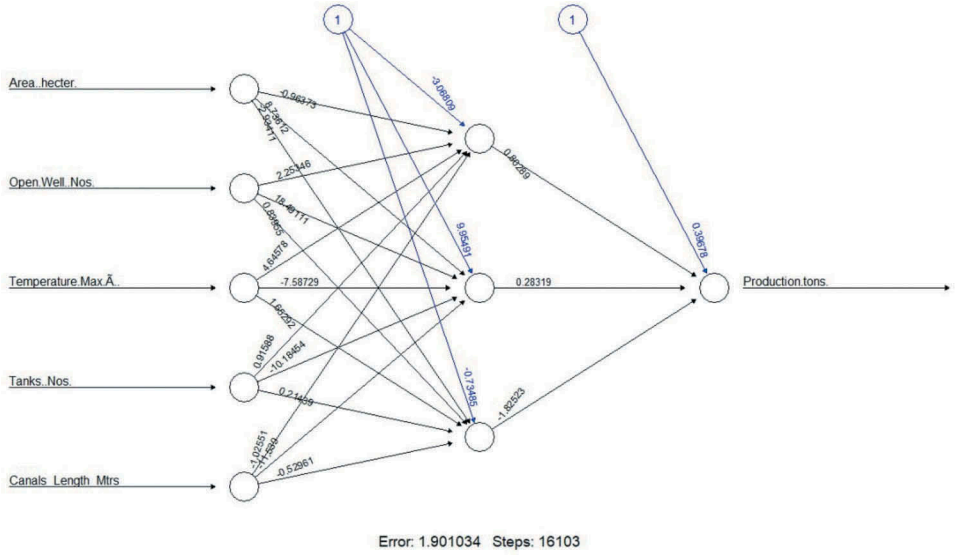


**Figure 5.** Neural network for FFS-based feature subset.

achieve better predictive performance (Nanda et al. 2018; Zhang and Huihua 2013). The tuning parameters of cost ($C$) and the kernel width ($\gamma$) are need to be set for RBF to obtain an accurate prediction. The $C$ and $\gamma$ value may vary in each and every feature subset. In order to obtain the optimal value of $\gamma$ and $C$ for each best feature subsets, SVR is tuned. The $C$ value set the range between 10 and 100 and $\gamma$ value set the range between 0 and 3, each step is increased by 0.1 to obtain

**Table 6.** The γ and C values for different feature subsets.

| Feature subset | $\gamma$ value | $C$ value |
|---|---|---|
| FFS | 0.3 | 15 |
| RF varImp | 0.2 | 25 |
| VIF | 0.1 | 10 |
| CBFS | 0.1 | 12 |

the optimal value. This procedure is applied to all distinct feature subsets. The $\gamma$ and $C$ values for different feature subsets are tabulated in Table 6.

## k-Nearest Neighbor

The KNN is a non-parametric approach (Denoeux, Kanjanatarakul, and Sriboonchitta 2015). The KNN approach is used for predicting the crop yield (Gonzalez-Sanchez, Frausto-Solis, and Ojeda-Bustamante 2014; Hansen and Indeje 2004; Shakil Ahamed et al. 2015). The KNN algorithm, k is the tuning parameter which plays a major role to obtain the accurate prediction. It finds a group of k sample (training data) that are nearest to unknown samples (test data). From these k samples, the unknown samples are determined by calculating the average of the response variable. The parameter of k is determined by using elbow criterion method. Figure 6 shows the relationship between elbow value and the number of clusters for different feature subsets. The optimal k value for all feature subsets obtained for the current work is 22.

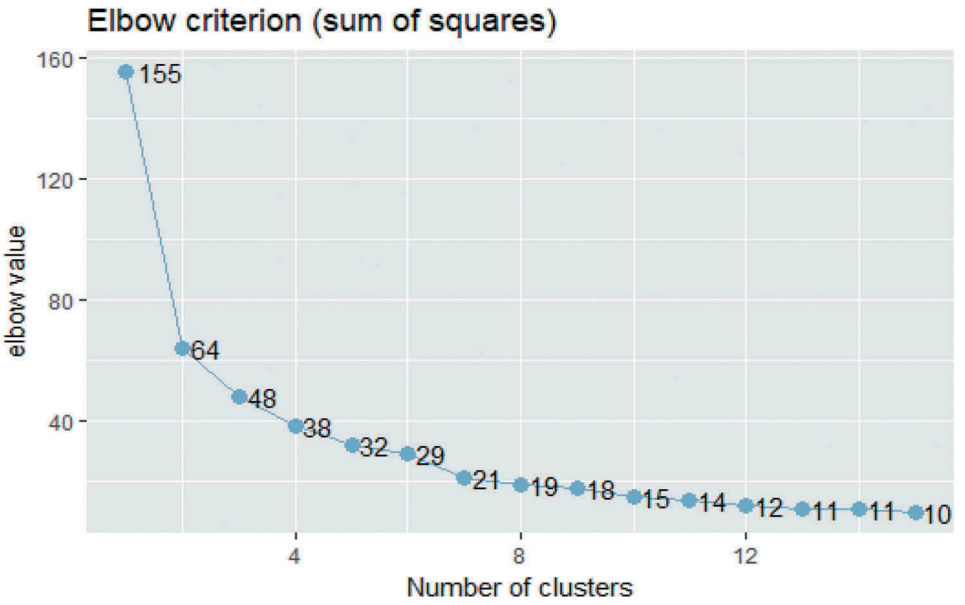


**Figure 6.** The relationship between elbow value and the number of clusters for different feature subsets.

### *Random Forest*

RF is an ensemble ML algorithm. RF is used to predict the crop yield (Dharumarajan and Rajendra Hegde 2017; Gonzalez-Sanchez, Frausto-Solis, and Ojeda-Bustamante 2014; Mathieu and Aires 2018). In RF, the prediction is based on averaging the randomized forests. In order to implement the RF, two parameters such as the number of trees (ntree) and the number of features in each split (mtry) have to setup. The accuracy of the prediction is based on these parameters. The parameters ntree and mtry may vary one feature subset to another feature subset. In order to obtain the optimal RF model for accurate prediction, ntree = 100, 500, and1000; mtry = 1:15 with a step size of 1 are fixed for this study and values for both parameters are tested and evaluated. The FFS feature subset achieved less RMSE value when the mtry = 3 and ntree = 1000. The CBFS feature subset achieved less RMSE value when mtry = 9 and ntree = 1000. The VIF feature subset achieved less RMSE value when mtry = 6 and ntree = 1000. The RFVarImp feature subset achieved less RMSE value when mtry = 12 and ntree = 1000. The different mtry and ntree values for distinct feature subsets of different algorithms show a similar pattern. The different mtry and ntree values for distinct feature subsets of RF algorithm are shown in Figure 7.

In addition to that the out-of-bag (OOB) error decreased sharply when ntree increased from 1 to 100. When ntree increased from 101 to 200, different feature subsets of the OOB error are slightly high and low in certain ntrees. When ntree increased from 200 to 400, the effect is same like 100 to 200. When ntree

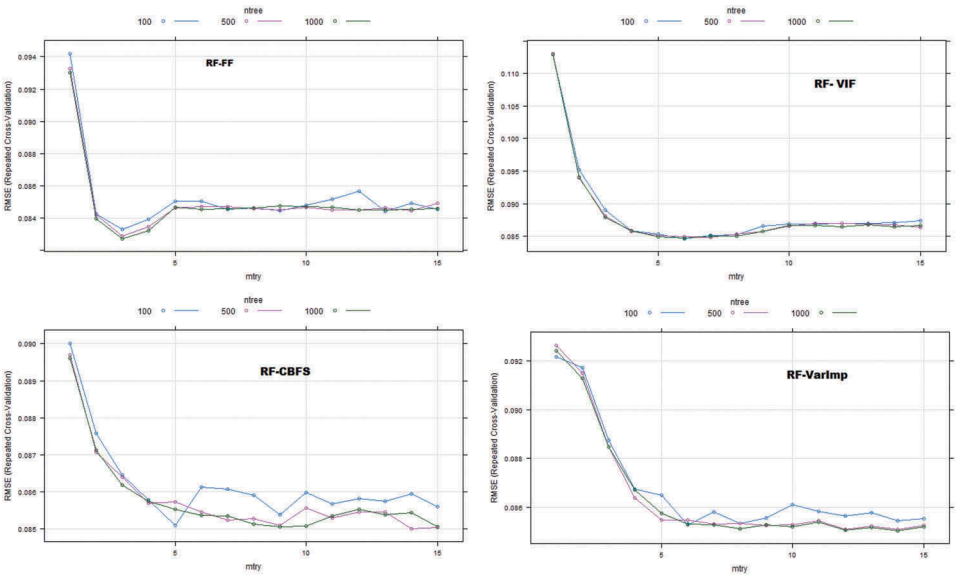


**Figure 7.** Effect of the number of trees and the number of random split variables at each node (mtry) of distinct feature subsets of RF algorithm.

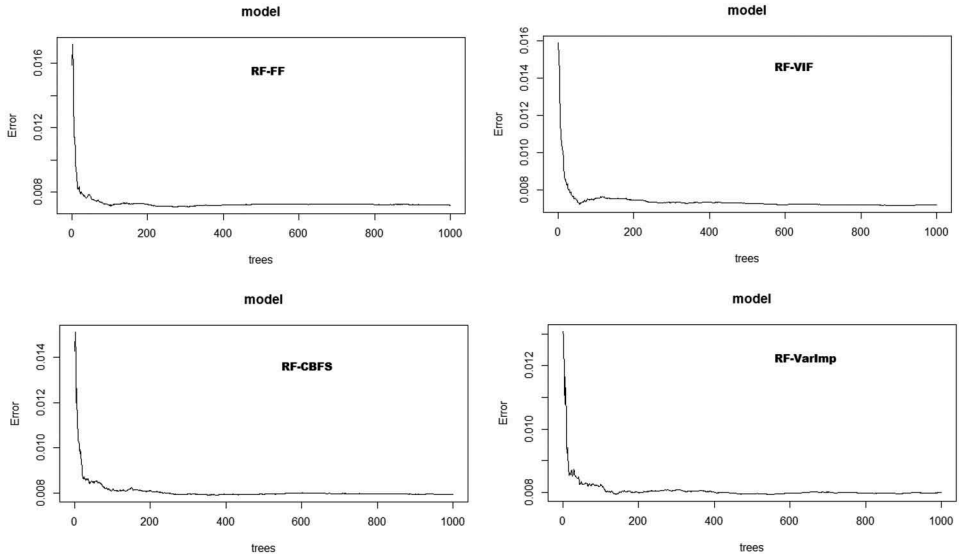


**Figure 8.** The relationship between out-of-bag (OOB) error and the number of trees of distinct feature subsets for random forest algorithm.

increased from 400 to 600 and 600 to 800, the trend is same like 100 to 200. When ntree increased from 800 to 1000, OOB error is decreased. Therefore, for all the distinct feature subsets the ntree value is 1000. The (OOB) error of distinct feature subsets for the RF algorithm is shown in Figure 8.

## *Performance Metrics*

In order to analyze the performance of the algorithms, three of the most commonly used accuracy metrics of regression models were used: RMSE, correlation coefficient ($R$), and the relative mean absolute error (MAE) (Hand, Mannila, and Smyth 2001). The RMSE measures the difference between the actual and estimates, exaggerating the presence of outliers (Han and Kamber 2006). The researchers used RMSE as one of the important parameters to analyze the performance of CYP models (Liu, Goering, and Tian 2001). The correlation coefficient ($R$) is also included, which measures the linear relationship between regression model predictions and the real values. MAE is the average of differences in estimations (in physical units).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n}}$$

**Table 7.** RMSE value of different machine learning algorithm with feature subset.

| Feature subset algorithm \ Machine learning algorithm | ANN | SVR | KNN | RF |
|---|---|---|---|---|
| FFS | 0.098 | 0.099 | 0.127 | 0.085 |
| CBFS | 0.104 | 0.118 | 0.098 | 0.093 |
| VIF | 0.106 | 0.106 | 0.091 | 0.088 |
| RFVarImp | 0.102 | 0.098 | 0.082 | 0.093 |

**Table 8.** MAE value of different machine learning algorithm with different feature subsets.

| Feature subset algorithm \ Machine learning algorithm | ANN | SVR | KNN | RF |
|---|---|---|---|---|
| FFS | 0.064 | 0.065 | 0.089 | 0.055 |
| CBFS | 0.080 | 0.080 | 0.065 | 0.060 |
| VIF | 0.070 | 0.070 | 0.070 | 0.056 |
| RFVarImp | 0.063 | 0.063 | 0.061 | 0.060 |

**Table 9.** *R*-value of different machine learning algorithm with different feature subsets.

| Feature subset algorithm \ Machine learning algorithm | ANN | SVR | KNN | RF |
|---|---|---|---|---|
| FFS | 0.92 | 0.92 | 0.87 | 0.93 |
| CBFS | 0.91 | 0.88 | 0.92 | 0.92 |
| VIF | 0.91 | 0.90 | 0.94 | 0.94 |
| RFVarImp | 0.91 | 0.92 | 0.94 | 0.93 |

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

$$MAE = \left(\frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{(n)(\bar{y})}\right)$$

The current work is the first study reporting and comparing the performance of different ML algorithms in combination with different feature selection algorithms. Each of the ML algorithms have provided varying degree of prediction accuracy as per the features selected using the feature selection algorithm. Prediction accuracy of different ML with different feature sets selected by different feature selection algorithms is calculated and analyzed by RMSE, MAE, and *R* performance metrics. Tables 7, 8, and 9 shows the results of RMSE, MAE, and *R* metrices for all evaluated techniques in different features‘ subset.

### *Performance of the Predictive Algorithms*

Figure 9a–d shows the scatterplot of predicted versus observed value of the model that is trained and evaluated for dependent variable for different predictive algorithms with the input features selected by different feature

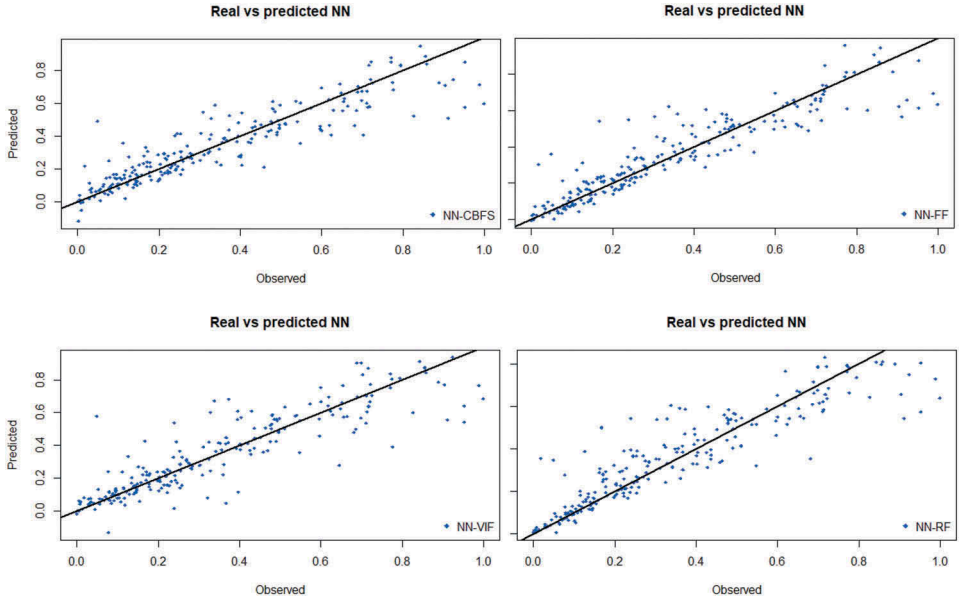


**Figure 9a.** The scatterplot of predicted versus the observed value of ANN.

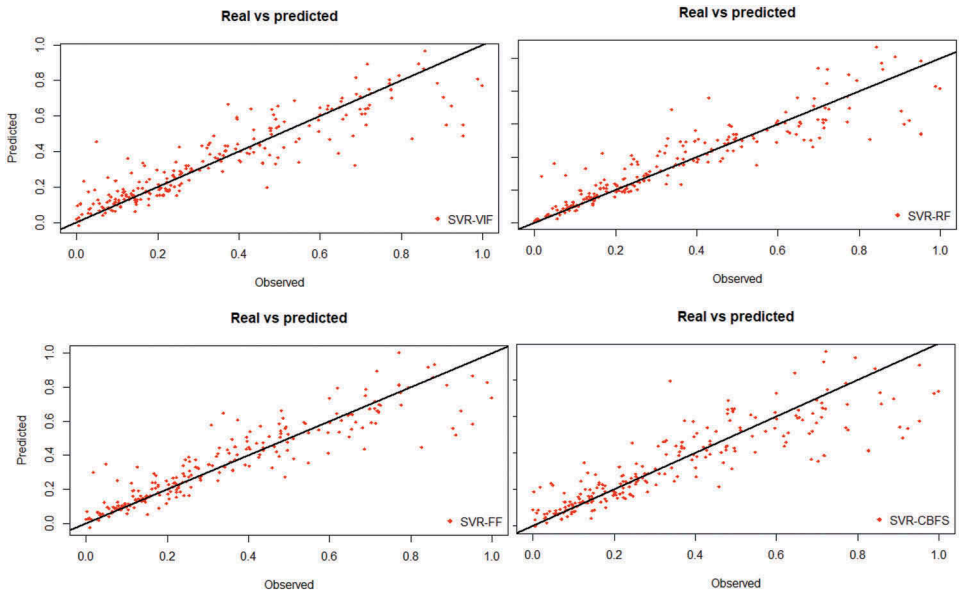


**Figure 9b.** The scatterplot of predicted versus the observed value of the SVR.

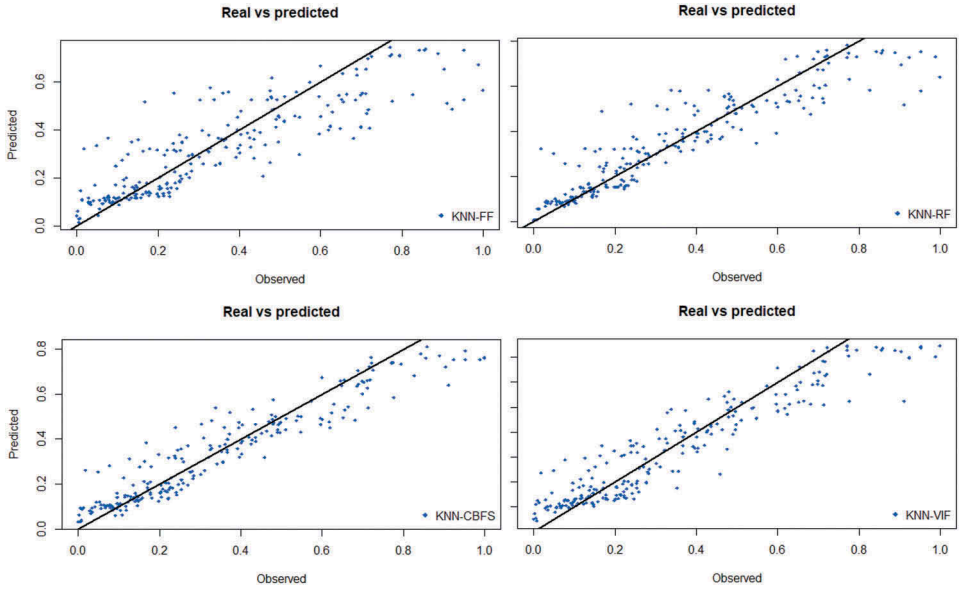


**Figure 9c.** The scatterplot of predicted versus the observed value of KNN.

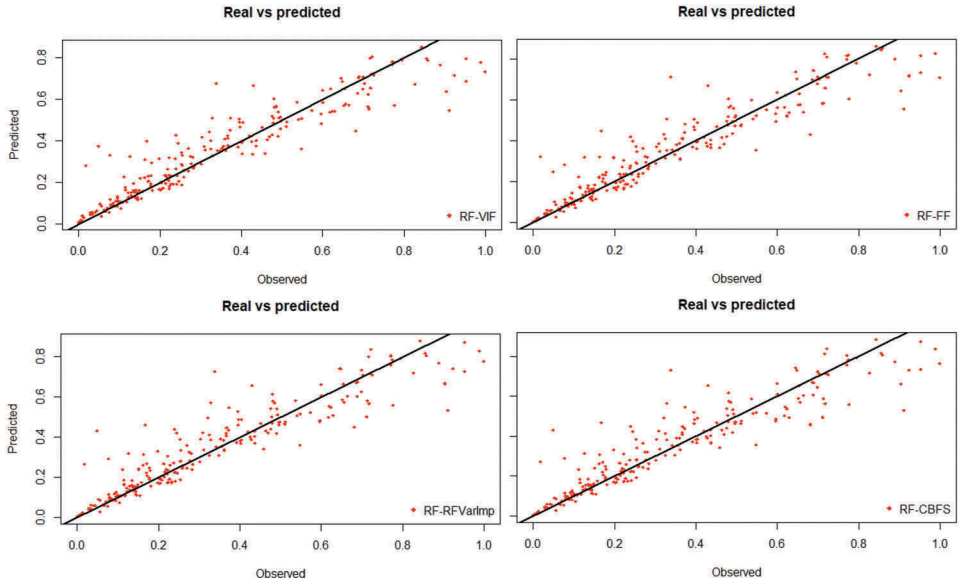


**Figure 9d.** The scatterplot of predicted versus the observed value of RF.

selection algorithms. It is demonstrated that the RF model input supplied with FFS models show more closeness between the actual and predicted values. It indicates that this model is good to predict the CYP. In all the other cases it clearly indicates that the outliers are more and the prediction is poor compared with RF with FFS. The RF with FFS handle the data well, and

the other models might be better suited to extrapolating from training data with short values to obtain accurate CYP.

## Discussion

As per the agricultural data, there is no benchmark data sets. Data are area sensitive. Therefore, it is very difficult to compare the performance accuracy with other data sets. From area to area the value of each variable may vary.

The predictive ability of ANN, SVR, KNN and RF algorithms with distinct feature subsets which are selected by FFS, CBFS, VIF, and RFVarImp selection algorithms are in Table 7, Table 8 and Table 9. The predictive ability is calculated by RMSE, MAE, and *R* metrices. Considering the RMSE value of different predictive algorithms with distinct feature subset, the RF algorithm headed the top position. FFS with RF has 0.085, ANN has 0.098, SVR has 0.099 and KNN has 0.127. Similarly, CBFS with RF has 0.093, KNN has 0.098, ANN has 0.104, and SVR has 0.118. The VIF with RF has 0.088, KNN has 0.091, ANN has 0.106, and SVR has 0.106. The RFVarImp with RF has 0.093, KNN has 0.082, SVR has 0.098, and ANN has 0.102. Based on overall performance of all the algorithms RF is the good predictive algorithm. Considering distinct feature subsets FFS gives good accuracy.

Considering the MAE value of different predictive algorithms with distinct feature subset, the RF algorithm headed the top position. FFS with RF has 0.055, ANN has 0.064, SVR has 0.065, and KNN has 0.089. Similarly, CBFS with RF has 0.060, KNN has 0.065, ANN has 0.080, and SVR has 0.080. The VIF with RF has 0.056, KNN has 0.070, ANN has 0.070, and SVR has 0.070. The RFVarImp with RF has 0.060, KNN has 0.061, SVR has 0.063, and ANN has 0.063. Based on overall performance of all the algorithms, RF is the good predictive algorithm. Considering distinct feature subsets FFS gives good accuracy.

The *R*-value of different predictive algorithms with distinct feature subset, the RF algorithm headed the top position. FFS with RF has 0.93, ANN has 0.92, SVR has 0.92, and KNN has 0.87. Similarly, CBFS with RF has 0.92, KNN has 0.92, ANN has 0.91, and SVR has 0.88. The VIFFS with RF has 0.94, KNN has 0.94, ANN has 0.91, and SVR has 0.90. The RFVarImp with RF has 0.94, KNN has 0.94, SVR has 0.92, and ANN has 0.91. Based on overall performance of all the algorithms RF is the good predictive algorithm. Considering distinct feature subsets FFS gives good accuracy.

It is observed that the RF algorithm works well in all the four distinct feature subsets. If the feature subset is highly correlated with the dependent variable and less correlation with the independent variable and based on the importance of each features KNN is more sensitive next to RF.

According to the results of this study, a random forest is a useful algorithm for prediction of crop yield using the input parameters related to the features

like area in hectare, number of open wells, number of tube wells, canal length and maximum temperature.

## Conclusion

The predictive ability of ANN, SVR, KNN and RF algorithm with distinct feature subsets which are selected by FFS, CBFS, VIF, and RFVarImp are evaluated and compared with each other. The features which are selected by FFS with RF algorithm produces the highest accuracy. It is concluded that the features such as production area, canal length, open well, tanks and maximum temperature which are selected by FFS algorithm give better accuracy when it is applied with RF algorithm.

## Acknowledgments

The authors gratefully acknowledge the Statistical Department and Agricultural Department of the State Government of Tamil Nadu and the Meteorological Department in India for providing the statistical data for the agrarian purpose.

## References

Akbar, A., A. Kuanar, J. Patnaik, A. Mishra, and S. Nayak. 2018. Application of Artificial Neural Network modeling for optimization and prediction of essential oil yield in turmeric (Curcuma longa L.). *Computers and Electronics in Agriculture* 148:160–78.

Bijanzadeh, E., Y. Emam, and E. Ebrahimie. 2010. Determining the most important features contributing to wheat grain yield using supervised feature selection model. *Australian Journal of Crop Science* 4 (6):402–07.

Breiman, L. 2001 Aug. Statistical modeling: The two cultures. *Statistical Science* 16 (3):199–215.

Brown, J. N., Z. Hochman, D. Holzworth, and H. Horan. 2018. Seasonal climate forecasts provide more definitive and accurate crop yield predictions. *Agricultural and Forest Meteorology* 260–261:247–54.

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

Denoeux, T., O. Kanjanatarakul, and S. Sriboonchitta. 2015. EK-NNclus: A clustering procedure based on the evidential K-nearest neighbor rule. *Knowledge-Based Systems* 88:57–69.

Dharumarajan, S., and S. K. Rajendra Hegde. 2017. Singh, Spatial prediction of major soil properties using Random Forest techniques – A case study in semi-arid tropics of South India. *Geoderma Regional* 10:154–62.

Drummond, S. T., K. A. Sudduth, A. Joshi, S. J. Birrel, and N. R. Kitchen. 2003. Statistical and neural methods for site-specific yield prediction. *T Asabe* 46 (1):5–14.

Fortin, J. G., F. Anctil, L. Parent, and M. A. Bolinder. 2011. Site specific early season potato yield forecast by neural network in Eastern Canada. *Precision Agriculture* 12:905–23.

Ghodsi, R., R. Mirabdollah Yani, R. Jalali, and M. Ruzbahman. 2012. Predicting wheat production in iran using an artificial neural networks approach. *International Journal of Academic Research in Business and Social Sciences* 2:2. February 2012.

Gonzalez-Sanchez, A., J. Frausto-Solis, and W. Ojeda-Bustamante. 2014. Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research* 12 (2):313–28.

Green, T. R., J. D. Salas, and R. H. E. Ana Martinez. 2007. Relating crop yield to topographic attributes using spatial analysis neural networks and regression. *Geoderma* 139 (2007):23–37.

Gu, Y. H., S. J. Yoo, C. J. Park, Y. H. Kim, S. K. Park, J. S. Kim, and J. H. Lim. 2016. BLITE-SVR: New forecasting model for late blight on potato using support-vector regression. *Computers and Electronics in Agriculture* 130:169–76.

Han, J., and M. Kamber. 2006. *Data mining: Concepts and techniques*. 2nd ed. USA: Morgan Kaufmann Publications.

Hand, D., H. Mannila, and P. Smyth. 2001. *Principles of data mining*. London, England: MIT Press.

Hansen, J. W., and M. Indeje. 2004. Linking dynamic seasonal climate forecasts with crop simulation for maize yield prediction in semi-arid Kenya. *Agricultural and Forest Meteorology* 125:143–57.

Hearst, M. A., B. Schölkopf, S. Dumais, E. Osuna, and J. Platt. 1998. Trends and controversies-support vector machines. *IEEE Intelligent Systems* 13:18–28.

Henryson, K., C. Sundberg, T. Kätterer, and P.-A. Hansson. 2018. Accounting for long-term soil fertility effects when assessing the climate impact of crop cultivation. *Agricultural Systems* 164:185–92.

Hill, D. J. 2010. B.S. Minsker Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software* 25:1014–22.

Ji, B., Y. Sun, S., . Yang, and J. Wan. 2007. Artificial neural network for rice yield prediction in mountainous regions. *Journal of Agricultural Science* (2007 (145):249–61.

Johnson, M. D., W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédar. 2016. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology* 218–219:74–84.

Liu, J., C. E. Goering, and L. Tian. 2001. A neural network for setting target corn yields. *Transactions of the ASAE* 44:705–13.

Mathieu, A., and F. Aires. 2018. Assessment of the agro-climatic indices to improve crop yield forecasting Jordane. *Agricultural and Forest Meteorology* 253–254:15–30.

Mehnatkesh, A., A. Sh., A. Jalalian, and A. A. Dehghani. Prediction of rainfed wheat Grain yield and biomass using artificial neural networks and multiple linear regressions and determination the most factors by sensitivity analysis, information technology, automation and precision farming. In *International Conference of Agricultural Engineering - CIGR-AgEng 2012: Agriculture and Engineering for a Healthier Life*. CIGR-EurAgEng Valencia, Spain. 2012 8-12 July 1554. ref.9.

Mobarake, S. A., M. Almassi, A. Hemmat, and M. Z. RezaMoghaddasi. 2014. A model for the estimation of yield and investigation on factors affecting irrigated wheat production in various tillage methods (using artificial neural networks). *Bulletin of Environment, Pharmacology and Life Sciences* 3 (5):79–84.

Monisha Kaul, R. L., and C. W. Hill. 2005. Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems* 85:1–18.

Nanda, M. A., K. B. Seminar, D. Nandika, and A. Maddu. 2018. A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information* 9:5.

Nari, K., and L. Yang-Won. 2016. Machine learning approaches to corn yield estimation using satellite images and climate data: A case of IOWA state. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography* 34 (4):383–90.

Oreski, D., S. Oreskib, and B. Klicek. 2017. Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing* 52:109–19.

Ruß, G., (2009). Data mining of agricultural yield data: A comparison of regression models. Proc. 9th Indust. Conf. on Advances in Data Mining-Applications and Theoretical Aspects, July 20-22, Leipzig, Germany.

Satir, O., and S. Berberoglu. 2016. Crop yield prediction under soil salinity using satellite derivedvegetation indices. *Field Crops Research* 192:134–43.

Shakil Ahamed, A. T. M., N. T. Mahmood, N. Hossain, M. T. Kabir, K. Das, F. Rahman, and R. M. Rahman, Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh, IEEE SNPD 2015, June 1-3 2015, Takamatsu, Japan.

Shamseldin, A. Y. 1997. Application of a neural network technique to rainfall-runoff modeling. *Journal of Hydrology* 199:272–94.

Shine, P., M. D. Murphy, J. Upton, and T. Scully. 2018. Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms. *Computers and Electronics in Agriculture* 150:74–87.

Taylor, C., B. Cullen, M. D'Occhio, L. Rickards, and R. Eckard. 2018. Trends in wheat yields under representative climate futures: Implications for climate adaptation. *Agricultural Systems* 164:1–10.

Torkashvand, A. M., A. Ahmadi, and N. L. Nikravesh. 2017. Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network (ANN) and multiple linear regressions (MLR). *Journal of Integrative Agriculture* 16 (7):1634–44.

Ximena, C., S. Rivera, J. Bacenetti, A. Fusi, and M. Niero. 2017 15 August. The influence of fertiliser and pesticide emissions model on life cycle assessment of agricultural products: The case of Danish and Italian barley. *Science of the Total Environment* 592:745–57.

Ying-Xue, S., X. Huan, and Y. Li-Jiao. 2017. Support vector machine-based open crop model, (SBOCM): Case of rice production in China. *Saudi Journal of Biological Sciences* 24:537–47.

Zhang, G., and G. Huihua. 2013. Support vector machine with a Pearson VII function kernel for discriminating halophilic and non-halophilic proteins. *Computational Biology and Chemistry* 46:16–22.