International Journal of Agricultural and Environmental Information Systems
Volume 10 • Issue 3 • July-September 2019

Selection of Important Features for Optimizing Crop Yield Prediction

Maya Gopal P S, VIT University, Chennai, India
Bhargavi R, School of Computing Science and Engineering, VIT University, Chennai, India

ABSTRACT

In agriculture, crop yield prediction is critical. Crop yield depends on various features including geographic, climate and biological. This research article discusses five Feature Selection (FS) algorithms namely Sequential Forward FS, Sequential Backward Elimination FS, Correlation based FS, Random Forest Variable Importance and the Variance Inflation Factor algorithm for feature selection. Data used for the analysis was drawn from secondary sources of the Tamil Nadu state Agriculture Department for a period of 30 years. 75% of data was used for training and 25% data was used for testing. The performance of the feature selection algorithms are evaluated by Multiple Linear Regression. RMSE, MAE, R and RRMSE metrics are calculated for the feature selection algorithms. The adjusted R2 was used to find the optimum feature subset. Also, the time complexity of the algorithms was considered for the computation. The selected features are applied to Multilinear regression, Artificial Neural Network and M5Prime. MLR gives 85% of accuracy by using the features which are selected by SFFS algorithm.

KEYWORDS

Artificial Neural Network, Feature Selection Algorithm, M5Prime, Model Validation, Multiple Linear Regression, Performance Metrics

1. INTRODUCTION & RELATED WORK

Data mining is a process of discovering previously unknown and potentially interesting patterns in large datasets (Frawley et al., 1991). The data mining process includes fixing the problem, understanding the data, preparing the data, applying the right techniques to build the models, interpreting the results and use the data into action. Now-a-days, intelligent data mining and knowledge discovery by artificial neural network and feature selection algorithms have become the important revolutionary concepts in prediction and modelling (Roddick et al., 2001, Schuize et al., 2005). Data set may contain redundant information that does not directly impact the predictions, and it may contain highly correlated attributes. The data sets are typically not gaining any new information by including all the attributes. In data mining, feature selection algorithms are useful for identifying irrelevant attributes to be excluded from the dataset (Che et al., 2017, Kotu et al., 2015). Feature selection in predictive analytics refers to the process of identifying few most important features or attributes that are essential in building a model for an accurate prediction. Efficient predictive models can improve the quality of the decision making. Feature selection optimizes the performance of the data mining algorithm and makes it easier for the analyst to interpret the outcome of the modeling. This procedure can reduce not only the cost of recognition by reducing the number of features to be collected, but in some cases it can also provide a better classification of prediction accuracy due to finite sample size effects (Jain et

DOI: 10.4018/IJAEIS.2019070104

al., 1982). This strategy aims at further reducing the number of features. Adding feature selection to the analytical process has several benefits: it simplifies and narrows down the scope of the features that are essential in building a predictive model, to minimize the computational time and memory requirements using the feature selection algorithms (Pal and Foody, 2010). The focus can be directed to a subset of predictors which are very essential.

Researchers work with different feature selection models to optimize their data sets. Automated feature selection for every algorithm with the conventional approach of stepwise regression for feature selection (Alvarez, 2009). Gonzalez- Sanchez et al. (2014), performed an exhaustive search of the feature selection algorithms. H. Liu et al. (1996) proposed a consistency based feature selection mechanism to evaluate the worth of a subset of the attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes. In this model the consistency of any subset can never be lower than that of the full set of attributes. M. Hall (1999) proposed a correlation based approach to feature selection in different datasets and demonstrated how it can be applied to both classification and regression problems for machine learning. Karimi et al. (2013) presented a hybrid feature selection methods by combining symmetric uncertainty measure and gain measure. Both measures for each feature-class correlation were calculated first and then rank feature according to average score value. High ranked feature greater than a threshold values was selected. They evaluated their system using knowledge discovery data dataset and Naïve Bayes algorithm. Correlation based method, Gain Ratio method and Information Gain method were used by Chaudhary et al. (2013) and presented the performance evaluation of three feature selection methods with optimized Naïve Bayes is performed on mobile device. Zhang et al. (2005) performed a principal components analysis to transform data and used stepwise feature selection for multiple linear regression (MLR). In most experiments conducted, researchers collect data that are supposedly related to the phenomenon of interest, given resource and/or time constraints on the collection and analysis of data. The oriented collection of data means that these kinds of datasets have only pre-approved features. Feature selection can enhance model quality by discarding unwanted features or simply decreasing the model and computational complexity by keeping the most important features with an example (Ruß et al., 2010).

The major research on agricultural management is the development of cost effective methods for predicting the crop yield with the available parameters like irrigation details, fertiliser details and temperature (Bocca et al., 2016; Zeynep Özcan et al., 2017). In recent years, the data analysts are working towards developing predictive models in the form of expert systems to improvise agricultural yield, considering the environmental features, soil quality, irrigation and land usage (Aggarwal, 1995; Bagley et al., 2012; De wit et al., 1987; Safa et al., 2015; Patricio Grassinia et al., 2015). The crop yield is classified into long term (years ahead), medium term (months to years ahead) and short term (weeks to months ahead). Long term paddy crop is focuses on this reserach work. Recent research focused on applying the machine learning and data mining techniques to predict the yield (Majumdar et al., 2017). Several researchers used regression models for crop yield prediction to analyse its applicablity with other models (Zhang et al., 2005; Ji et al., 2007; Alvarez et al., 2009; Gonzalez-Sanchez et al., 2014; Matsumura et al., 2014). Researchers compared the classical statistical models against Artificial Neural Networks, regression trees and support vector regression for better crop yield models (Drummond et al., 2003; Fortin et al., 2011; Ruß, 2010; Felipe et al., 2016).

The aim of this research work is to identify important paddy field conditions (features) using feature selection algorithms for providing a comprehensive view about paddy crop yield. Understanding the importance of features among a large dataset of features can play a key role in improving the yield under field conditions. Our study investigates the behaviour of five feature selection algorithms with sixteen features and the outcome is given as input to multiple linear regression model, artifical neural network and M5Prime to find the accuracy.

2. DATA SOURCE

The cumulative agricultural related data of thirty years is collected from the Statistical Department, Meteorological Department and Agricultural Department. The statistical data with land cover area for the study area (Tamilnadu State, India) has been gathered from the Department of Economics and Statistics, Government of Tamil Nadu. The latitude of Tamil Nadu, India is 11.127123, and the longitude is 78.656891. Tamil Nadu is located in India in the States place category with the GPS coordinates of 11°7' 37.6428'' N and 78°39' 24.8076'' E. Tamil Nadu, India elevation is 138 meters height that are equal to 453 feet. The statistical data along with the agricultural production data and weather data are the two sets of data collected for this work. The regular data collected are used as such from the departments. In order to achieve accurate prediction, the data are normalized between 0 and 1. Since the data set is multidimensional and the range of the value is large. The normalised data are combined into a single data set, run through feature selection algorithms and predictive model. It provides predictive value for paddy crop yield in the particular area. The analyzing area has an average climate. The agricultural production data contains planting area, irrigation area, fertilizer usage and irrigation details. The weather data set comprises climatological features including rainfall, maximum and minimum temperatures and solar radiation. Each instance of the dataset contains the details about the crop with a cultivated area, annual production and weather features monitored during the year. As a result, 745 instances with 16 features including area of cultivation (hectare), canal length (m), tanks (nos.), tube wells (nos.), open wells (nos.), production (tons), rain fall (mm), maximum temperature (°C), minimum temperature (°C), average temperature (°C), solar radiation (W/m^2), seed quantity (kg), nitrogen, phosphorus and potassium applied to the soil (kg) and yield (ton/hectare) are documented for 30 years upto 2017. The feature type, data type and feature category are listed in Table 1. An identification (ID) is given for each feature along with the description. Before applying the feature selection algorithm the data set is pre-processed for identification and understanding of features, missing values treatment and outlier treatment. The quality of data will decide the quality of the output.

3. DATA PRE-PROCESSING

Before applying the data into the feature selection algorithms, they are pre-processed for filling the missing values, outlier detection and data transformation. Missing value is treated by using a learning algorithm, missForest algorithm. The Boxplot and the Histogram graphical tools are used in checking the normality assumption and identifying potential outliers. In this work, values below Q1 – 1.5 IQR or above Q3 + 1.5 IQR is treated as an Outlier by the IQR (Inter Quartile Range) Rule. In this work, transform the data within a specific range using V in [min, max] to V' in [0,1], by applying V' = (V-Min)/(Max-Min).

4. MODELS AND METHODS

4.1. Feature Selection and Evaluation

Machine learning is a computational learning approach that focuses on prediction from statistical data. Feature selection algorithm is applied to recognize important features which are strong correlation with crop yield. This becomes even more important when the number of features are very large. The algorithms can be assisted by feeding in only those features that are relevant. Feature subsets give better results than complete set of feature for the same algorithm with less computational time. Major reasons to use feature selection are that it enables the machine learning algorithm to train faster, reduces the complexity of a model and makes it easier to interpret. It also improves the accuracy of a model if the right subset is chosen and reduces over fitting. The execution time of a particular algorithm is of much less importance than its ultimate classification performance for a moderate size

Table 1. Structure of the data set

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

feature sets. But for some recent applications have focused on performing feature selection on data sets with hundreds of features. In such cases, execution time becomes extremely important as it may be impractical to run some algorithms even once on such large data sets. In this case, feature selection plays an important role. Different statistical methods can be used in the feature selection such as filters, wrapper and embedded methods. The filter approach is a preprocessing step and, it does not consider the effects of a selected feature subset on the performance of the algorithm. Wrapper methods evaluate a subset of features according to accuracy of a given predictor. Embedded methods perform variable selection during the process of training and are generally specific to given algorithms. Sequential Forward Feature Selection (SFFS) algorithm & Sequential Backward Elimination Feature Selection (SBEFS) algorithm are wrapper methods, Correlation based Feature Selection (CBFS) algorithm & Variance Inflation Factor (VIF) algorithm are filter methods and Random Forest Variable Importance (RFVarImp) algorithm is embedded method. CBFS selects the best feature subset which is highly correlated with the production. VIF checks the multicollinearity between independent features. So it removes all multicollinear independent features. Random forest VarImp selects most important features based on information gained by individual features. SFFS and SBFS selects best feature and then selects the best including the best one. All the above mentioned five feature selection algorithms were used for selecting the features in this research work.

4.1.1. Sequencial Forward Feature Selection Algorithm

Forward feature selection is an iterative method that starts with null feature set and, for each step, the best feature that satisfies some criterion function is included with the current feature set, i.e., one step of the sequential forward selection is performed. The algorithm also verifies the possibility of improvement of the criterion if some feature is excluded. In this case, the worst feature as per the criterion is eliminated from the set, i.e., it performs one step of sequential backward selection. SFFS proceeds dynamically increasing and decreasing the number of features until the desired "n" features

are reached. The SFFS algorithm is based on Akaike Information Criterion (AIC) value for feature selection. The time complexity of the algorithm is O(n) for SFFS (Koller et al. 1996). This algorithm selects the AH, CL,TK, OW and Tmax features.The selection procedure is given in result section.

4.1.2. Sequencial Backward Elimination Feature Selection Algorithm

SBEFS involves starting with all candidate features, testing the deletion of each feature using a chosen model fit criterion, deleting and repeating this process until no further features can be deleted without a statistically significant loss of fit. The backward search starting with the entire feature set of size "m" and perform the search until the desired features "n" are reached. SFBS is based on the AIC value for feature selection and the time complexity is O(*n*). The backward search method requires more computation than forward selection process (Koller et al. 1996). This algorithm selects the AH, CL,TK, OW and Tmax features.The selection procedure is given in result section.

4.1.3. Correlation Based Feature Selection Algorithm

CBFS ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is towards subsets that contain features which have high correlation with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will have high correlation with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas that are not already predicted by other features. 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 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.

The bias of the evaluation function is towards subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have less correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. It requires m($(n^2-n)/2$) operations for computing the pair wise feature correlation matrix, where m is the number of instances and n is the initial number of features. The time complexity of the CBFS method is $O(2^n)$, where n is the number of features (L Yu et al. 2003). This algorithm selects AH, CL, TK, AT, Tmax, SD, NF, PF and KF.

4.1.4. Variance Inflation Factor

The Variance Inflation Factor quantifies the severity of multi-co-linearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of co-linearity. VIF method is used to remove correlated independant features. It is extremely fast and it uses 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 provably avoids model over-fitting. VIF is calculated by using the formula $VIF = \frac{1}{1-R_i^2}$. The time complexity is $O(n^2)$ (Lin et al. 2011), where n is the number of features. This algorithm selects AH, CL, TK, TW, OW, RainF, AT, Tmin, Tmax and SR features.

4.1.5. Random Forest Variable Importance

Random forests or random decision forests are an ensemble learning method for classification and regression.These operate by constructing a multitude of decision trees at training time and outputting

the class i.e., the mode of the classes or mean prediction of the individual trees. Random forests use a modified tree learning algorithm that selects, a random subset of the features, at each candidate split in the learning process. The time complexity of random forest is O(n*m log(m)), where m is the number of instances and n is the number of features (Witten et al. 2006).This algorithm selects the features AH, TK, OW, NF, PF, KF and SD.

4.2. Multiple Linear Regression Model

Multiple Linear Regression (MLR) is one of the statistical model that specifies how one set of features, called dependent features, functionally depend on another set of features, called independent features. It has been applied most frequently for crop yield prediction because production(yield) is normally dependent on number of parameters such as production area, irrigation,fertilizers, and weather parameters. Here production(yield) is the dependent feature and other parameters are independent features.

MLR model is descriped by

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{ik} + \varepsilon_i$$

where k is the number of feature x_{ij} is the i^{th} observation of the feature x_j,β_1, β_2, $\beta_{3,...}$ β_k are the regression coefficients and ε_i are the error term or residuals. The above equation can be writtern as $Y = \sum X\beta + \varepsilon$.

4.2.1. MLR Model for Crop yield Prediction

The regression equation defines the given regression model with six independent variables. Here PD is the dependent feature and AH, OW, Tmax, TK and CL are the independent features. The slope of variables indicates that, for a given value of the particular variable estimate to decrease or increase of predicted value.

$$PD = 0.005 + 0.961AH + 0.117OW + 0.049T\max - 0.078TK - 0.046CL$$

The scatter plot shown in figure 1 gives the linear relationship between two variables, the maximum temperature and production of crop. The increasing trend in the scatter plot indicates that the variables have a positive association.

Figure 2 shows the frequency distribution of the feature is based on a smooth curve. The horizontal axis from left to right indicates the different possible values of maximum temperature. The vertical axis from bottom to top measures the frequency of how many times a particular value occurs. Highest value for frequency will be at the top of the curve and lowest on both extremes.

4.3. M5 Prime

M5 Prime is one of the regression tree model which is used in crop yield prediction (Gonzalez-Sanchez et al. 2014).The tree construction method of M5 Prime is similar to CART(Classification and Regression Tree) but the regression trees are much smaller then CART. The M5 Prime(M5P) regression tree splits the samples' space recursively until regions are small enough to represented by a simple model (Quinlan, 1992). In M5 Prime, standard deviation is used as error criterion.

$$Error = \sigma(X) - \sum_i \frac{\left|X_i\right|}{\left|X\right|}\sigma(X_i)\ \sigma\left(X\right) - \sum_i \frac{Xi}{X}\sigma\left(Xi\right)$$

Figure 1. Temperature maximum vs production

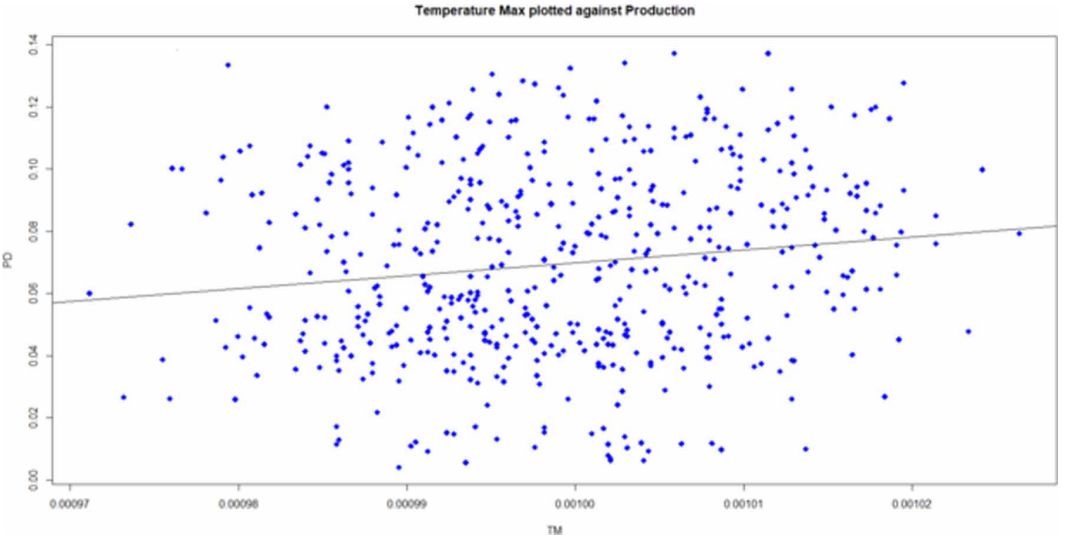
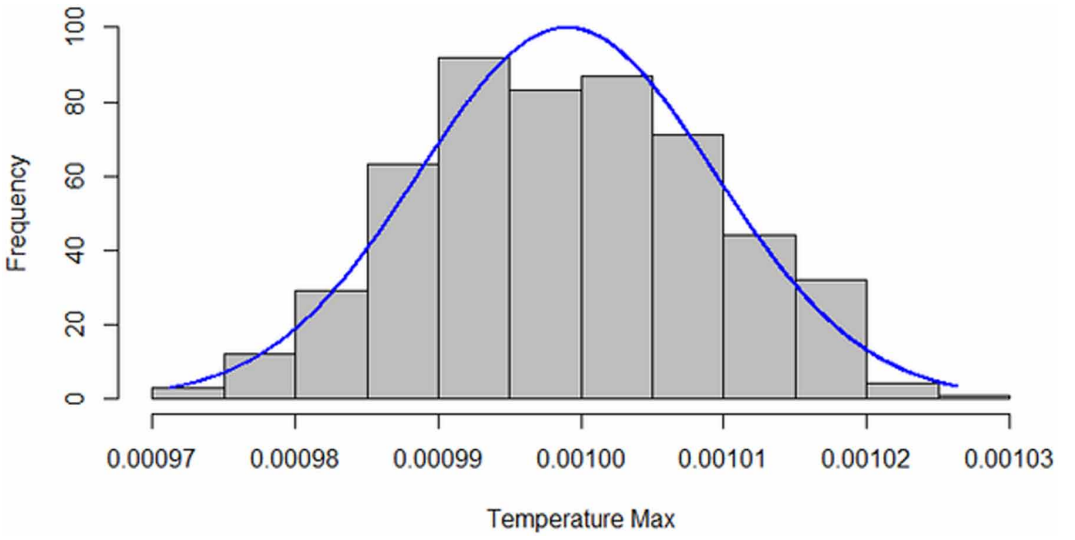


Figure 2. Frequency Distribution of temperature maximum



Where σ is standard deviation and i is the number of subregions of a region whose instances are denoted by X. After all possible splits, M5P selects the one with minimum error. M5P has some salient features than CART such as it can handle linear models at leaf nodes and a smoothing process is used in M5P for estimation of the response variable (Quinlan 1992). Because of these reasons M5P overcomes CART algorithm in accuracy.

4.4. Artificial Neural Network(ANN)

ANN is a supervised machine learning algorithm which is most commonly used in crop yield prediction (Safa et al. 2015]. It mimicis human nervous system. It has three layers: input layer, hidden layer and output layer. These layers are interconnected together via neurons of each layer. The costellation of

neurons and connenctions are called as architecture of the network. Each neuron receives the weighted activation of the other neurons through its incoming connections and then these are summed. The result is passed through an activation function. The outcome is the activation of the neuron. For each of the outgoing connections, this activation value is multiplied with the specific weight and transferred to the another neuron. The sigmod activation function is used. The input layer contains number of neurons which are equlivalent to number of input features such as AH,CL,TK,OW and Tmax and output layer contains only one neuron which is PD.

4.5. Accuracy Metrics

Four accuracy metrics such as root mean square error (RMSE), the mean absolute error (MAE), root relative mean square error (RRMSE), and correlation coefficient (R) are used for calculating the accuracy. The RMSE measures the difference between the actual and estimates, exaggerating the presence of outliers (Han & Kamber 2006). The RRMSE, which compares the model prediction against the mean. For this metric, a value below 100% indicates a better performance than the average. Thus, RRMSE is easy to read by people unaccustomed to crop yield dimensions. Correlation coefficient (R) measures the linear relationship between regression model predictions and the real values. MAE is the average of differences in estimations (in physical units). In addition to that R^2 value is calculated to check the accuracy of the model once the features are fitted into the model. It also measures the variance explained by the model. The equation given below shows how these metrics are calculated, where y is the real production (yield) value, $\hat{y}$ represents the production (yield) estimation, *i* is the number of instances, $\overline{y}$ is the average of the real production(yield) values, and $\overline{\hat{y}}$ is the average of predictions

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

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

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

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

$$R_{adj}^2 = 1 - \left[\frac{\left(1 - R^2\right)\left(n-1\right)}{n-k-1}\right]$$

5. RESULTS AND DISCUSSIONS

The current paper presents an implementation of the algorithm for finding all the relevant features by using R tool (R Development Core Team 2010). R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Experimental results are presented in this section.

By applying the various feature selection algorithms, the different features are selected based on the selection criteria. The selected features are listed in table 2. In all the cases the area of cultivation gain more importance for the crop production (yield). Number of tanks also play a crucial role for the selected dataset when compared with other predictors. The forward and backward feature selection algorithms have given same set of features and similar performance. It selects same set of features, since both the algorithms are based on AIC value for feature selection. The AIC value is calculated by using the formula $AIC = N \ln\left(\frac{SS_{error}}{N}\right) + 2K$, here N is the number of observation and K is the number of paramter +1. The forward method is faster than its backward counterpart due to its low time complexity. This is to be expected, as the forward method starts with a null set and enlarges them. The backward algorithm starts with large subset and reduce to the required feature. It is computationally more expensive to achieve the criterion value due to the large subsets. The forward feature selection takes computational time is O(n) where n is the number of features selected by the algorithm which satisfies the lowest AIC value. While adding one more feature on this set, the AIC value is increased. So at this point the selection procedure is stopped. In the current calculations consider PD as the dependent feature. Initially the AIC value is calculated with a null set and the AIC value is -3643.92. In the next step, AIC value is calculated for all the independent features individually. The lowest value is -4533.46 for AH an independent feature and it is selected in step1. In the next step, the OW independent feature along with AH having lowest AIC value as -4587.72 are selected. In the next step, the TK independent feature with AH and OW having the lowest AIC value as -4600.22 is selected. In the next step, CL along with AH, OW, TK has the lowest AIC value as -4604.99. Hence, this feature is selected. In the next step, Tmax with AH, OW, TK,CL has lowest AIC value as -4607.05 and it is selected. In the next step, NF with AH, OW, TK, CL and Tmax has the lowest AIC value as -4609.14 and these features are selected. In the next step,when SR is added in the set the AIC value is increased, the AIC value of this set is -4608.9, so at the previous step the selection process is stopped. Using forward feature selection algorithm the independent features AH, OW, TK, CL and Tmax are selected for the PD dependent feature. The AIC value for the subsets are shown in table 3.

In the backward elimination feature selection algorithm, the AIC value is calculated for all independent features for the PD dependent feature.The AIC value is -4603.66.The algorithm separates the SD independent feature and the AIC value is calculated. AIC value is same as previous one, hence SD is removed from the feature set. Then independent feature KF is separated and AIC value is calculated. AIC value is same as previous one and KF is removed from the feature set. Similarly

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	√	√	√		√					√				
Backward Elimination	√	√	√		√					√				
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
{ }	-3643.92	
{AH}	-4533.46	↓
(AH, OW}	-4587.72	↓
{AH, OW, TK}	-4600.22	↓
{AH, OW, TK, CL}	-4604.99	↓
{AH, OW, TK, CL,Tmax}	-4607.05	↓
{AH, OW, TK, CL, Tmax, NF}	-4609.14	Stop
{AH, OW, TK, CL, Tmax, NF, SR}	-4608.9	↑

the algorithm separates independent feature PF and the AIC value is calculated which is same as previous one. Hence PF is removed from the feature set. Then independent feature Tmin separates from the feature set and AIC value is calculated which is further reduced as -4605.36. Then RainF independent feature is separated from the feature set and AIC value is calculated. It is further reduced as -4606.89. Then TW independent feature is separated from the feature set and the AIC value is calculated which is further reduced as -4608.4. Then independent feature AT is separated from the feature set and the AIC value is calculated which is further reduced as -4608.85. Then independent feature SR is separated from the feature set and the AIC value is calculated which is further reduced as -4609.14. Then independent feature NF is separated from the feature set and the AIC value is calculated which is increased as -4607.1. So NF is eliminated from the feature set. The independent feature such as CL, TK, OW, AH and Tmax are selected for the PD dependent feature. Both the forward and backward selection algorithms select the same set of features for further evaluation. But backward elimination algorithm takes more time than the forward algorithm due to its high time complexity. The AIC values are listed in table 4.

VIF based feature selection algorithm, checks the colinearity among the independent features. Among the independent features AH, SD, NF, PF and KF are collinear. So only one independent feature AH is selected from the feature set and all other features are removed from the feature set. The other features such as CL, TK, TW, OW, RainF, AT, Tmin, Tmax and SR are selected since the

Table 4. Backward elimination feature selection procedure by using AIC

Feature Subset	AIC value	Selection procedure
{ CL, TK, TW, OW, AH, RainF, AT, Tmin, Tmax, SR, NF, PF, KF, SD}	-4603.66	
{ CL, TK, TW, OW, AH, RainF, AT, Tmin, Tmax, SR, NF, PF, KF}	-4603.66	↓
{ CL, TK, TW, OW, AH, RainF, AT, Tmin, Tmax, SR, NF, PF}	-4603.66	↓
{ CL, TK, TW, OW, AH, RainF, AT, Tmin, Tmax, SR, NF }	-4603.66	↓
{ CL, TK, TW, OW, AH, RainF, AT, Tmax, SR, NF}	-4605.36	↓
{ CL, TK, TW, OW, AH, AT, Tmax, SR, NF }	-4606.89	↓
{ CL, TK, OW, AH,, AT, Tmax, SR, NF }	-4608.4	↓
{ CL, TK, OW, AH, Tmax, SR, NF }	-4608.85	↓
{ CL, TK, OW, AH, Tmax, NF }	-4609.14	Stop
{ CL, TK, OW, AH, Tmax}	-4607.1	↑

non-colinearity of the features. The algorithm checks the colinearity of the features individually and it requires more computational time. The calculated VIF values are listed in table 5.

In the correlation based feature selection algorithm, initially correlation matrix is generated. The dependent feature PD, all the possible independent feature subsets are generated and the score is calculated. The highest score subset selects as the final sets of features. The subset contains TK, OW, NF, PF, KF, SD, AH and Tmax features give highest score. Hence these features are selected It does an exhaustive search and it needs more computational time. The correlation matrix is shown in figure 3.

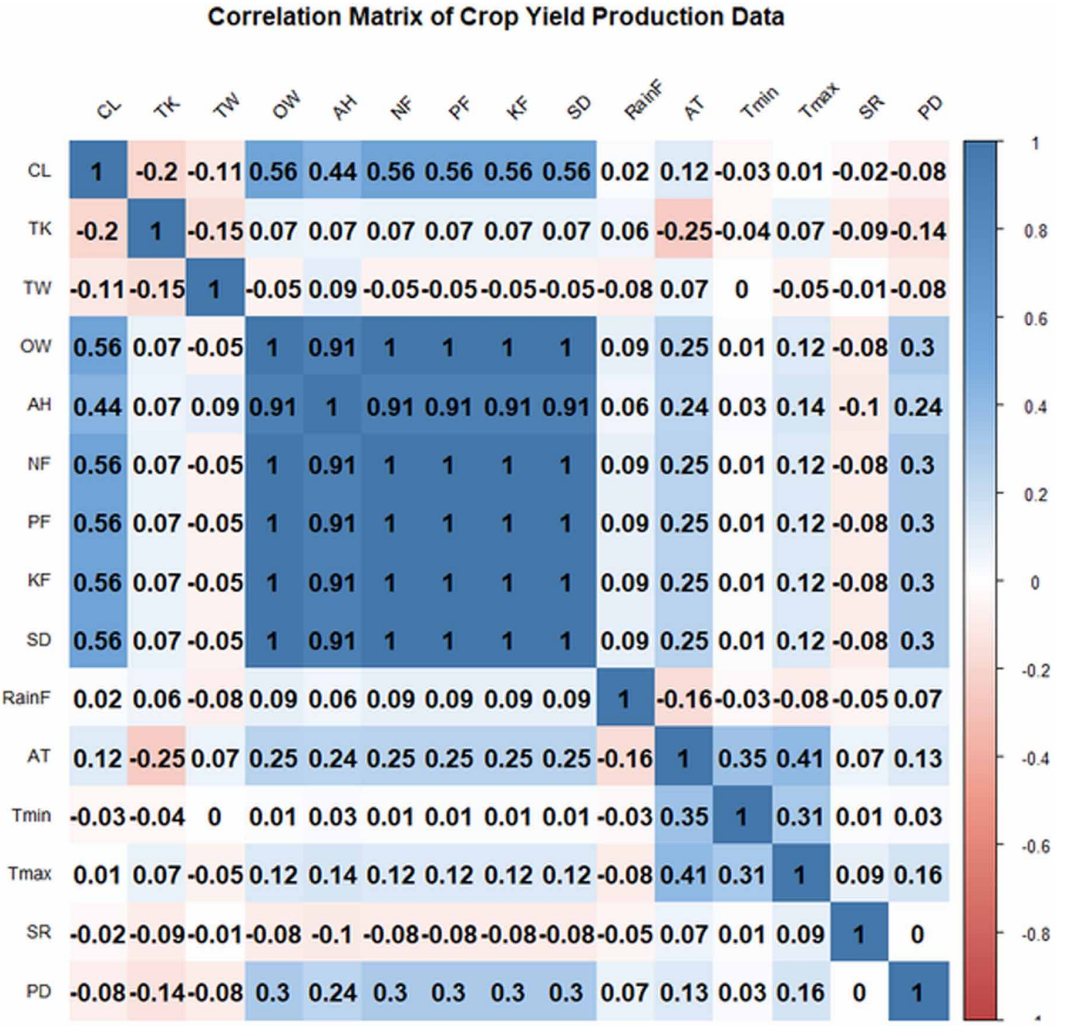
In the random forest variable importance, the features are selected based on the node purity. In this work, the node purity threshold is set as median of the node purity values. So the purity value above the median is selected. From the entire feature set TK, OW, AH, NF, PF, KF and SD are selected and the values are listed in table 6.

The features selected by applying the five feature selection algorithm are given in the table 2. Here each algorithm follows different selection criteria for selecting the features. To find the best features using feature selection algorithm with low time complexity the feature subset is given into the multiple linear regression model and find the fitness of the model based on the Adjusted R^2. The Adjusted R^2 showed the variance explained by the model. If the Adjusted R^2 is more, then that model is good. So that value is taken into consideration. In our work, forward feature selection and backward selection gives the same features and both gives same Adjusted R^2. But when the time complexity is considered forward feature selection is better than backward elimination algorithm. Based on the significance of the features given by model, the features AH, OW,TK,CL and Tmax is considered as the best features which are selected by forward feature selection algorithm. In this work, when no feature seletion is done and all features are given the input to the model, it gives less adjusted R^2 . Feature selection algorithms play a major role for better prediction. Once specific features are selected, it will take less time and space complexity. It leads the better prediction. Table 7 shows the RMSE, MAE, *R*, RRMSE and adjusted R2 measures using all the potential features when the features are given as inputs of MLR, ANN and M5P. The results show that SFFS features are give less error than other features which are given by other algorithms. The figure 4 shows the bar graph for the performance of feature selection algorithms based on the adjusted R2 values.

Table 5. Features and its VIF value

Feature	VIF values
CL	1.400091
TK	1.988810
TW	1.437965
OW	1.084505
RainF	1.058834
AT	1.591005
Tmin	1.201348
Tmax	1.336783
SR	1.028063
AH	2.236249

Figure 3. Correlation matrix of crop yield production data



6. MODEL VALIDATION

Figure 5 and figure 6 present the residual plot and fitted values versus residuel respectively for the regression model. The model is validated and the residual plot of the model shows that it does not follow any pattern. This model worked well and gives an accuracy of 85%. The current result for the feature selection algorithm applied MLR reach the commonly recommended accuracy of 85% (Anderson et al. 1976). This is the indication for the validity of the models studied.

Figure 7 shows the values for maximum temp versus residuals. In the multiple regression models checking the linearity assumption is not so straightforward. The direct way to illustrate the linearity is to plot the standardized residuals against each of the predictor variables in the regression model. The residual versus predictor variable plot in the figure indicates the random scatter of points.

Table 6. Features and its node purity based on Random Forest Variable Importance

Feature	IncNodePurity
CL	0.007241060
TK	0.017494514
TW	0.005762905
OW	0.014855902
AH	0.082870685
RainF	0.006314565
AT	0.009467588
Tmin	0.003521096
Tmax	0.003475371
SR	0.003464572
NF	0.079764258
PF	0.079882515
KF	0.077607548
SD	0.079586579

Figure 4. Performance of feature selection algorithms

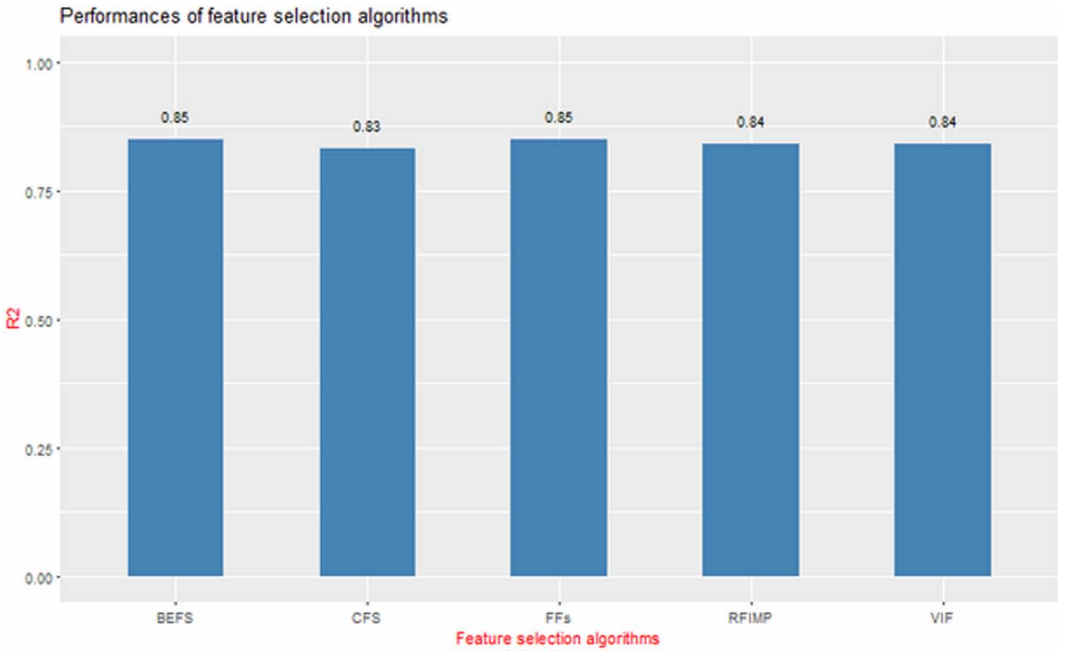
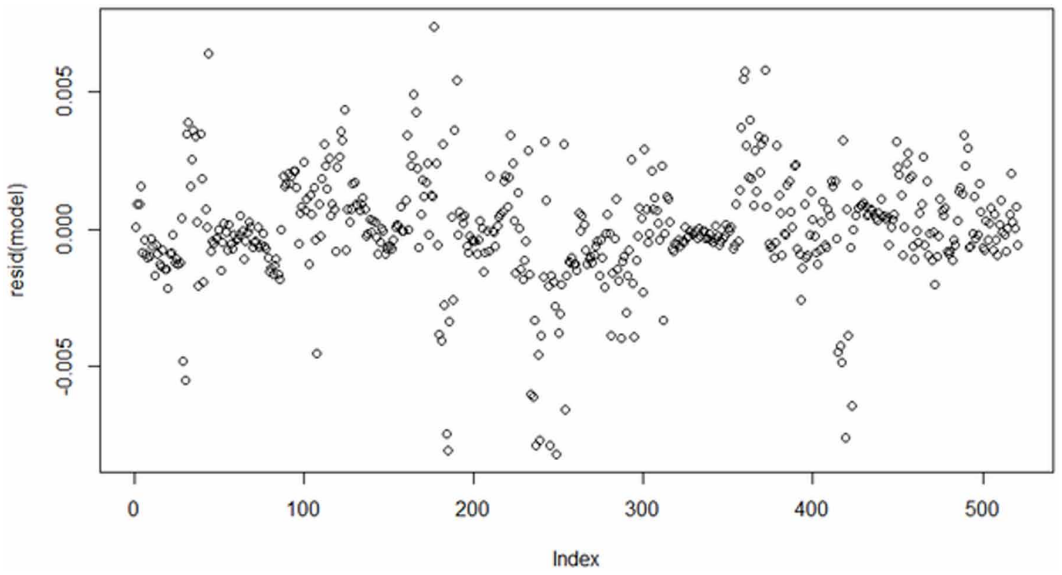


Table 7. Performance of the various feature selection algorithms

Metrics	FFS			BEFS			VIF			CFS			RFIMP		
	MLR	M5P	ANN	MLR	M5P	ANN	MLR	M5P	ANN	MLR	M5P	ANN	MLR	M5P	ANN
RMSE	0.013	0.078	0.098	0.013	0.078	0.098	0.013	0.085	0.106	0.014	0.085	0.104	0.014	0.081	0.102
MAE	0.009	0.054	0.064	0.009	0.054	0.064	0.009	0.058	0.070	0.009	0.056	0.080	0.009	0.055	0.063
R	0.898	0.94	0.92	0.898	0.94	0.92	0.898	0.93	0.91	0.888	0.93	0.91	0.895	0.93	0.91
RRMSE	0.012	0. 09	0.078	0.012	0. 09	0.078	0.012	0.010	0.08	0.013	0.011	0.081	0.012	0.012	0.08
Adj R^2	0.85	-	-	0.85	-	-	0.84	-	-	0.83	-	-	0.84	-	-

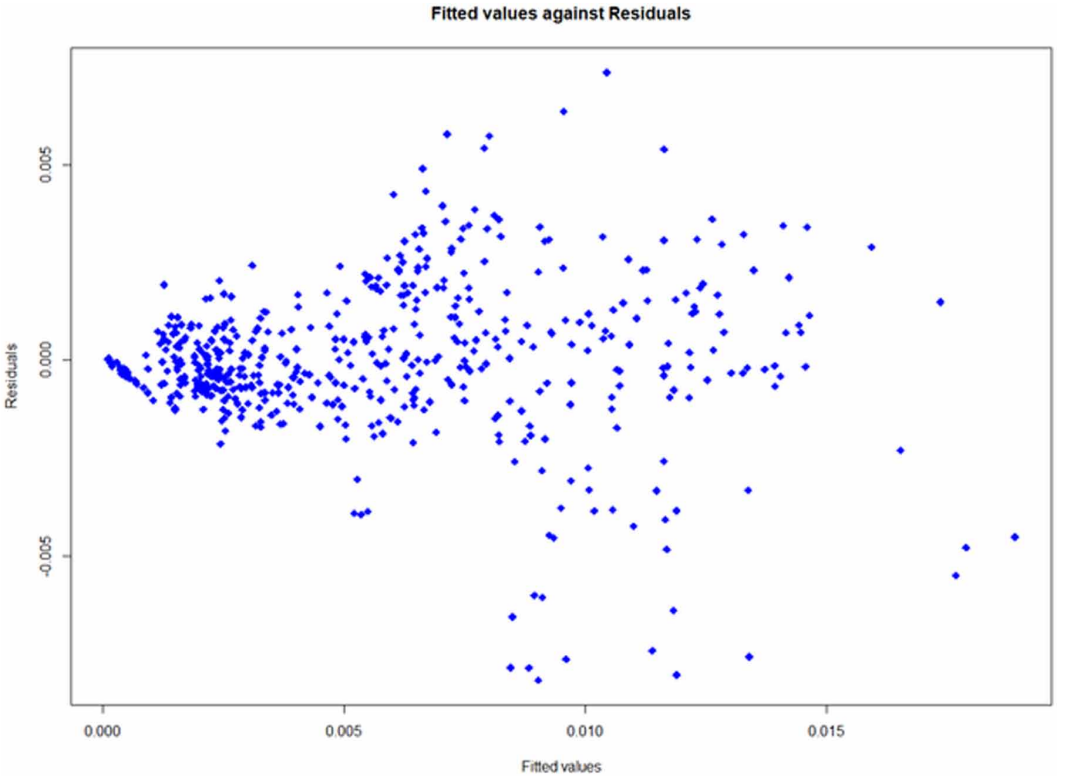
Figure 5. Residual plot



7. CONCLUSION

In this research work, five feature selection algorithms namely sequential forward feature selection, sequential backward feature elimination, correlation based feature selection, random forest Variable Importance and Variance Inflation Factor are applied in order to select the features. The feature selection algorithms select important features based on their selection criterian. RMSE, MAE, R and RRMSE metrics are calculated and used to analyse the performance of feature selection algorithms. The outcome of the feature selection is given to MLR, ANN and M5P. The MLR model accuracy is calculated using Adjusted R^2. Performance of all the algorithms gives small deviations in the RMSE, RRMSE and MAE values. 85% model accuracy achieved by using the features which are selected by forward selection algorithm and backward elimination feature selection algorithm when it applied to MLR. But considering the computational time forward feature selection takes less time. When all the features are given into the model gives 84% accuracy. It is concluded that, the forward feature selection algorithm is good for the better prediction and Area, Open Wells, Tanks, Temperature maximum and Canal Length are considered best features for the paddy dataset for given study area.

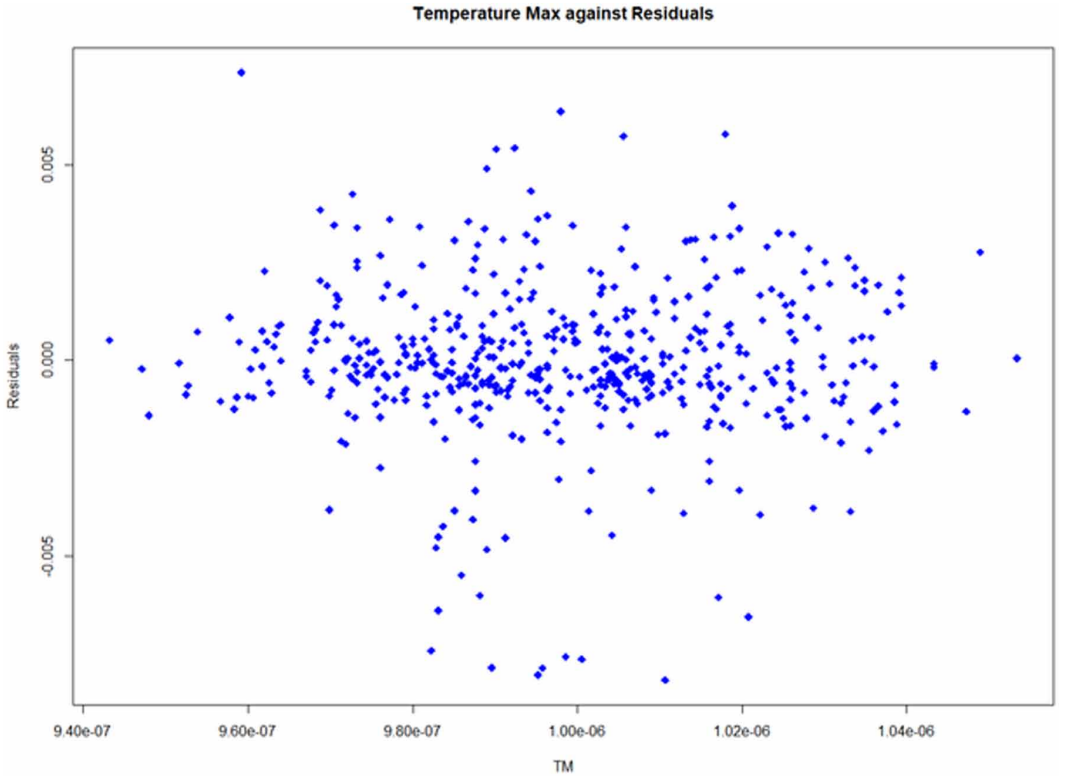
Figure 6. Fitted values versus residual



ACKNOWLEDGMENT

The authors gratefully acknowledge the Department of Economics and Statistics for providing the statistical data, Department of Agriculture Government of Tamil Nadu, India for providing the agricultural data and The Regional Meteorological Departmenent, Chennai for providing the weather data.

Figure 7. Temperature maximum versus residuals



REFERENCES

Aggarwal, P. K. (1995). Uncertainties in crop, soil and weather inputs used in growth models-implications for simulated outputs and their applications. *Agricultural Systems*, *48*(3), 361–384. doi:10.1016/0308-521X(94)00018-M

Alvarez, R. (2009). Predicting average regional yield and production of wheat in the argentine pampas by an artificial neural network approach. *European Journal of Agronomy*, *30*(2), 70–77. doi:10.1016/j.eja.2008.07.005

Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). *A Land Use and Land Cover Classification System for Use with Remote Sensor Data. Government Printing Office*. Washington, DC: US Geological Survey.

Bagley, J. E., Desai, A. R., Dirmeyer, P. A., & Foley, J. A. (2012). Effects of land cover change on moisture availability and potential crop yield in the world's breadbaskets. *Environmental Research Letters*, *7*(1), 1–9. doi:10.1088/1748-9326/7/1/014009

Choudhary, A., & Kolhe, S. (2013). Performance Evaluation of feature selection methods for Mobile devices. *Int. Journal of Engineering Research and Applications*, *3*(6), 587–594.

De Wit, C. T., & Van Keulen, H. (1987). Modelling production of field crops and its requirements. *Geoderma*, *40*(3-4), 253–265. doi:10.1016/0016-7061(87)90036-X

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

Bocca, F. F., & Luiz, H. A. R. (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. doi:10.1016/j.compag.2016.08.015

Fortin, J. G., Anctil, F., Parent, L., & Bolinder, M. A. (2011). Sitespecific early season potato yield forecast by neural network in Eastern Canada. *Precision Agriculture*, *12*(6), 905–923. doi:10.1007/s11119-011-9233-6

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1991). Knowledge Discovery in Databases: An Overview. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.), *Knowledge Discovery in Databases*. Cambridge, MA: MIT Press.

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, *12*(2), 313. doi:10.5424/sjar/2014122-4439

Jain, A. K., & Chandrasekaran, B. (1982). Dimensionality and Sample Size Considerations. In P. R. Krishnaiah & L. N. Kanal (Eds.), *Pattern Recognition in Practice* (Vol. 2, pp. 835–855). North- Holland.

Majumdar, J., Naraseeyappa, S., & Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: Application of big data. *J Big Data*, *4*(1), 20. doi:10.1186/s40537-017-0077-4

Che, J., Yang, Y., Li, L., Bai, X., Zhang, S., & Deng, C. (2017). Maximum relevance minimum common redundancy feature selection for nonlinear data. *Information Sciences*, *409–410*, 68–86. doi:10.1016/j.ins.2017.05.013

Ji, B., Sun, Y., Yang, S., & Wan, J. (2007). Artificial neural networks for rice yield prediction in mountainous regions. *J. Agr. Sci.*, *145*(03), 249–261. doi:10.1017/S0021859606006691

Hall, M. (1999). *Feature Selection for Discrete and Numeric Class Machine Learning*. Department of Computer Science, The University of Waikato.

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques* (2nd ed.). Morgan Kaufmann Publ.

Karimi, Z., Mansour, M., & Harounabadi, A. (2013, September). Feature Ranking in Intrusion Detection Dataset using combination of filtering. *International Journal of Computers and Applications*, *78*.

Koller, D., & Sahami, M. (1996). Toward optimal feature selection. In *Proceedings 13th International Conference on Machine Learning*, Bari, Italy. San Mateo, CA: Morgan Kaufmann.

Lin, D., Foster, D. P., & Ungar, L. H. (2011). VIF regression: A fast regression algorithm for large data. *Journal of the American Statistical Association*, *106*(493), 232–247. doi:10.1198/jasa.2011.tm10113

Liu, H., & Setiono, R. (1996, July). A probabilistic approach to feature selection-a filter solution. In the 13th International Conference on Machine Learning ICML (Vol. 96, pp. 319-327).

Matsumura, K., Gaitan, C.F., Sugimoto, K., Cannon, A.J., & Hsieh, W.W. (2014). Maize yield forecasting by linear regression and artificial neural networks in Jilin. *China. J. Agr. Sci. FirstView*, 1–12.

Grassinia, P., van Busselb, L. G. J., Van Wart, J., Wolf, J., Claessens, L., Yang, H., & Kenneth, G. et al. (2015). Cassman " How good is good enough? Data requirements for reliable crop yieldsimulations and yield-gap analysis. *Field Crops Research*, *177*, 49–63. doi:10.1016/j.fcr.2015.03.004

Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(5), 2297–2307. doi:10.1109/TGRS.2009.2039484

Quinlan, J. R. (1996). Learning with continuous classes. In *Proceedings AI92* (pp. 343–348).

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roddick, J. F., Hornsby, K., & Spiliopoulou, M. (2001). An updated bibliography of temporal, spatial, and spatio-temporal data mining research. In Temporal, Spatial, and Spatio-Temporal Data Mining (pp. 147-163). Springer. doi:10.1007/3-540-45244-3

Ruß, G., & Kruse, R. (2010). Feature selection for wheat yield prediction. In M. Bramer et al. (Eds.), *Research and development in intelligent systems XXVI*. London: Springer-Verlag. doi:10.1007/978-1-84882-983-1_36

Safa M., Samarasinghe S., & Nejat M., (2015). Prediction of wheat production using artificial neural networks and investigating indirect factors affecting it: Case study in Canterbury Province, New Zealand. *J. Agr. Sci. Tech.*, *17*, 791-803.

Schuize, F. H., Wolf, H., Jansen, H., & Vander, V. P. (2005). Applications of artificial neural networks in integrated water management: Fiction or future? *Water Science and Technology*, *52*(9), 21–31. doi:10.2166/wst.2005.0279 PMID:16445170

Kotu, V., & Deshpande, B. (2014). Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann.

Witten, I. H., & Frank, E. (2006). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann Publishers.

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 856-863).

Zhang, B., Valentine, I., & Kemp, P. (2005). Modelling the productivity of naturalised pasture in the north island, New Zealand: A decision tree approach. *Ecological Modelling*, *186*(3), 299–311. doi:10.1016/j.ecolmodel.2004.12.016

Özcan, Z., Kentel, E., & Alp, E. (2017). Evaluation of the best management practices in a semi-arid region with high agricultural activity. *Agricultural Water Management*, *194*, 160–171. doi:10.1016/j.agwat.2017.09.007

Mayagopal P.S is currently pursuing Doctoral Degree in the School of Computing Science and Engineering, VIT University, Chennai Campus, India. She has more than 15 years of Industry, Academic and Research experience. She received her M.E from Anna University, Chennai. Her research interests include Data Mining, Machine learning and Data Science.

Bhargavi R (Bhargavi Rentachintala is presently working as Associate Professor in the School of Computing Science and Engineering, VIT University, Chennai Campus, India. She has more than 20 years of Industry, Academic and Research experience She received her M.Tech and Ph.D degrees from IIT Madras and Anna University respectively. Her research interests include Complex Event Processing, Machine learning, in Healthcare and Data Science. She has authored and published several research papers in IEEE/ACM/Springer international conferences and refereed Journals. She also authored chapters in highly reputed research reference books.