Journal of Agrometeorology 23 (2) : 242-248 (June 2021)

Prediction for rice yield using data mining approach in Ranga Reddy district of Telangana, India

BABY AKULA*, R.S.PARMAR[1], M. P. RAJ[1] and K. INDUDHAR REDDY[2]

CoA, PJTSAU, Rajendranagar, Hyderabad, India
[1]CoAIT, Anand Agricultural University, Anand, India
[2]ARC, PJTSAU, Rajendranagar, Hyderabad, India
*Corresponding author: *babys_akula@yahoo.co.in*

ABSTRACT

In order to explore the possibility of crop estimation, data mining approach being multidisciplinary was followed. The district of Ranga Reddy, Telangana State, India has been chosen for the study and its year wise average yield data of rice and daily weather over a period of 31 years i.e. from 1988-2019 (30th to 47th Standard Meteorological Weeks). Data mining tool WEKA (V3.8.1). Min- Max Normalization technique followed by Feature Selection algorithm, 'cfsSubsetEval' was also adopted to improve quality and accuracy of data mining algorithms. Thus, after cleaning and sorting of data, five classifiers *viz.,* Logistic, MLP (Multi Layer Perceptron), J48 Classifier, LMT (Logistic Model Trees) and PART Classifier were employed over the trained data. The results indicated that the function based and tree based models have better performance over rule based model. In case of function based two models examined, *viz*., Logistic and MLP, the later performed better over Logistic model. Between tree based two models, LMT performed better over J48. Thus, MLP classifier model found to be the best fit model in predicting rice yields as it recorded an accuracy of 74.19 %, sensitivity of 0.742 and precision of 0.743 as compared with other models. The MLP has also achieved the highest F1 score of (0.742) and MCC (0.581).

Key words: Data mining, rice yield prediction, logistic, multi layer perceptron, logistic model

Reliable prediction of crop yield plays a vital role in agricultural management in India as being agriculture based economy. Various classification techniques such as the Naive Bayes, J48, random forests, support vector machines, artificial neural networks were used for crop yield prediction (Sujatha *et al*., 2016). Basically, three approaches are followed to study reliable yield prediction based on visual observations, weather and other input parameters and on biometrical characters of growing crop. Data mining is one of the important drivers for agriculture development being multidisciplinary as it merges artificial intelligence, computer science, machine learning, database management, mathematics algorithms and statistics (Liao, 2003).

Data mining is the process of analyzing, extracting and predicting the meaningful information from huge data to mine different pattern (Jain, 2009). A data mining methodology in agriculture was proposed by Fathima and Geetha (2014). Classifiers such as K-means clustering and kNN classifiers used had achieved 76% of accuracy in identifying suitable land, crop variety, etc. to get high yield. Mucherino *et al*., (2009) followed data mining techniques that include K-means, K nearest neighbor, ANN (Artificial Neural Network), and support vector machines in predicting the yields.

Classification algorithms such as J48, REPTree, and Random Forest have been applied over agricultural data set for predicting crop productivity with prediction accuracy of 83% (Diriba and Borena, 2013). Rani and Vidyavathi (2013) while predicting sugarcane yield using PCA and decision tree classifier revealed that Random Forest, Decision Stump, REP Tree and J48 achieved accuracies of 75%, 100%, 100%, and 99%, respectively. Similarly, Singh *et. al*., (2017); Biswas and Singh (2020) reported that climatic variables explained variation in yield by 28 per cent using Random Forest test in north Punjab. Thus, in the light of these findings, the present study was devised with the main objective of examining the influence of weather parameters in estimation of rice crop yield in Ranga Reddy district of Telangana using data mining approach.

MATERIALS AND METHODS

In order to explore the possibility of predicting effect of weather on rice yields in Ranga Reddy district of Telangana open source Data mining tool WEKA version 3.8.1 was used. Necessary rice yield data was collected from Directorate of Economics and Statistics, Government of Telangana, India and weather data from Agro-climate Research Centre, PJTS

Table. 1: Details of variables included in week-wise approach

S.No.	BSS	Rainfall	Max. temp.	Min. temp.	Morning (RH_1)	Evening (RH_2)
1	Z_{130}	Z_{230}	Z_{330}	Z_{430}	Z_{530}	Z_{630}
2	Z_{131}	Z_{231}	Z_{331}	Z_{431}	Z_{531}	Z_{631}
3	Z_{132}	Z_{232}	Z_{332}	Z_{432}	Z_{532}	Z_{632}
4	Z_{133}	Z_{233}	Z_{333}	Z_{433}	Z_{533}	Z_{633}
5	Z_{134}	Z_{234}	Z_{334}	Z_{434}	Z_{534}	Z_{634}
6	Z_{135}	Z_{235}	Z_{335}	Z_{435}	Z_{535}	Z_{635}
7	Z_{136}	Z_{236}	Z_{336}	Z_{436}	Z_{536}	Z_{636}
8	Z_{137}	Z_{237}	Z_{337}	Z_{437}	Z_{537}	Z_{637}
9	Z_{138}	Z_{238}	Z_{338}	Z_{438}	Z_{538}	Z_{638}
10	Z_{139}	Z_{239}	Z_{339}	Z_{439}	Z_{539}	Z_{639}
11	Z_{140}	Z_{240}	Z_{340}	Z_{440}	Z_{540}	Z_{640}
12	Z_{141}	Z_{241}	Z_{341}	Z_{441}	Z_{541}	Z_{641}
13	Z_{142}	Z_{242}	Z_{342}	Z_{442}	Z_{542}	Z_{642}
14	Z_{143}	Z_{243}	Z_{343}	Z_{443}	Z_{543}	Z_{643}
15	Z_{144}	Z_{244}	Z_{344}	Z_{444}	Z_{544}	Z_{644}
16	Z_{145}	Z_{245}	Z_{345}	Z_{445}	Z_{545}	Z_{645}
17	Z_{146}	Z_{246}	Z_{346}	Z_{446}	Z_{546}	Z_{646}
18	Z_{147}	Z_{247}	Z_{347}	Z_{447}	Z_{547}	Z_{647}

Where, Z_{1i} = Weekly Average of BSS for ith week; Z_{2i} = Weekly Total Rainfall for ith week; Z_{3i} = Weekly Average of Max.T. for ith week; Z_{4i} = Weekly Average of Min.T. for ith week; Z_{5i} = Weekly Average of RH_1 for ith week; Z_{6i} = Weekly Average of RH_2 for ith week; (i=30, 31, 32, 33, 34,…., 47 Meteorological Standard Week)

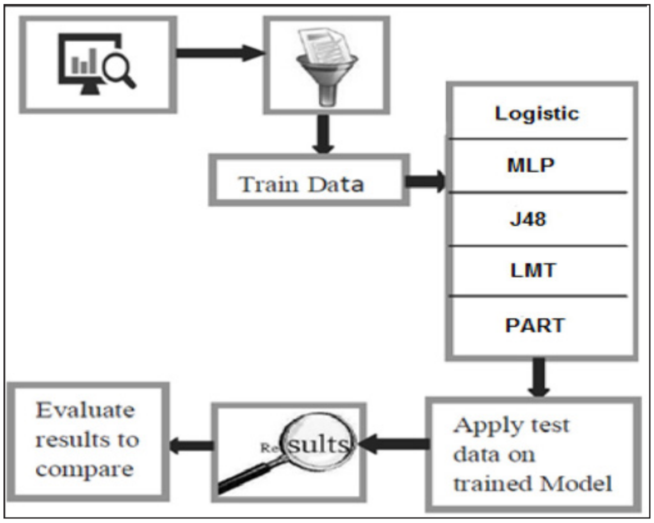
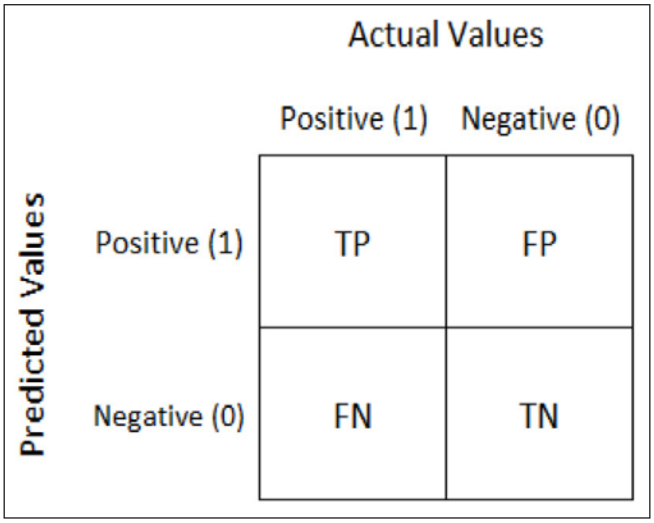


Fig. 1: Architecture of the rice yield prediction system



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 2: Confusion matrix

Agriculture University, Hyderabad, India

The district wise average yield data of rice and daily weather data were used over a period of 31 years *(*1988-2019) for 30th to 47th Standard Meteorological Weeks). Thus, total attributes of weekly averaged data of bright sunshine hours, temperature (maximum and inimu), relative humidity (morning and evening) and weekly total rainfall were used to judge combined effect on rice yields and are listed in Table 1.

(Total attributes = 18 * 6 = 108, i.e. Z_{130} ,Z_{131},…… Z_{646},Z_{647})

Steps followed while predicting rice yields

1. Initially, rice yields were categorized as L - Low, N - Normal and H – High. Dataset was prepared in Excel sheet with CSV extension for analysis by open source data mining tool WEKA (V3.8.1). **Min-Max** Normalization technique was used to normalize the experimental dataset

Table 2: Comparison of the results for each classification models

Parameters	Logistic	Multilayer perceptron (MLP)	J48	LMT	PART
Correctly classified instances (%) (Prediction Accuracy)	67.74	74 .19	64.51	67.74	61.29
Incorrectly classified instances (%)	32.26	26.81	35.48	32.26	38.71
Kappa Statistics	0.49	0.59	0.45	0.49	0.38
Mean Absolute Error (MAE)	0.22	0.21	0.28	0.28	0.29
Root Mean Squared Error (RMSE)	0.46	0.39	0.47	0.41	0.49
Relative Absolute Error (RAE) (%)	51	48	66	65	68
Root Relative Squared Error (%)	99	83	102	88	105

which reduces large variation of yield prediction. Feature Selection algorithm *viz*., **'cfsSubsetEval'** was also used as it improves the quality of the data and increases the accuracy of data mining algorithms. It also focuses on eliminating redundant and irrelevant data (Fig. 1).

2. The raw data cleaned and sorted.
3. The classifiers *viz*., Logistic, MLP, J48, LMT and PART were then employed over the trained data.
4. The results of each algorithm were noted from WEKA and compared with each other.
5. Correctly Classified Instances, Incorrectly Classified Instances, Receiver Operating Characteristic (ROC) Area, Precision Recall Curve (PRC) Area, Kappa Statistic, Mean Absolute Error, Relative Absolute Error, Root Mean Squared Error and Root Relative Squared Error, True Positive Rate, False Positive Rate, F-Measure, Recall, Precision and MCC values were taken into consideration for each case.
6. Thereafter performance was measured using three factors namely Sensitivity, Specificity, and Accuracy.
7. General Definitions :
   - TP (True Positive) = Number of cases where classification model detects for a condition when it is really present.
   - TN (True Negative) = Number of cases where classification model does not detect a condition when it is not present.
   - FN (False Negative) = Number of cases where classification model does not detect a condition when actually it is present.
   - FP (False Positive) = Number of cases where classification model detects a condition when it is really not present.
8. Performance evaluation factors used for performance measurement were as follows:
   - Confusion Matrix- summarizes the performance of a classification model (Fig.2)
   - RMSE- measures the difference between the predicted values and the actual values
   - MAE- measures the difference between two continuous variables
   - RAE - gives the total absolute error between the variables
   - Sensitivity-defined as percentage of correctly classified cases. It is True Positive Rate (TPR)
   - Specificity-defined as percentage of incorrectly classified cases. It is True Negative Rate (TNR)
   - Accuracy-defined as the overall success rate of a classifier.
   - F1 score- measure of test's accuracy and value ranging from 0 (worst) to 1 (best).
   - MCC- measure of quality of the classification.
   - Kappa Statistics- frequently used to test interrater reliability.
   - ROC Area- performance measurement for classification problem at various thresholds settings.

RESULTS AND DISCUSSION

The open source and free tool weka was used for knowledge analysis. WEKA has a set of classification models. The results are revealed with 10 fold cross validation to avoid overlapping. Form 108 independent attributes, 11 sig-

Table 3: Comparison of the statistics of the classification models

Parameters	Classifier model				
	Function based		Tree based		Rule based
	Logistic	Multilayer perceptron(MLP)	J48	LMT	PART
Correctly Classified Instances (Accuracy) (%)	67.74	74 .19	64.51	67.74	61.29
TP Rate (Sensitivity)	0.677	0.742	0.645	0.677	0.613
FP Rate (Specificity)	0.200	0.170	0.185	0.172	0.232
Precision	0.713	0.743	0.650	0.673	0.579
F1 score	0.688	0.742	0.647	0.674	0.591
MCC	0.495	0.581	0.455	0.503	0.378

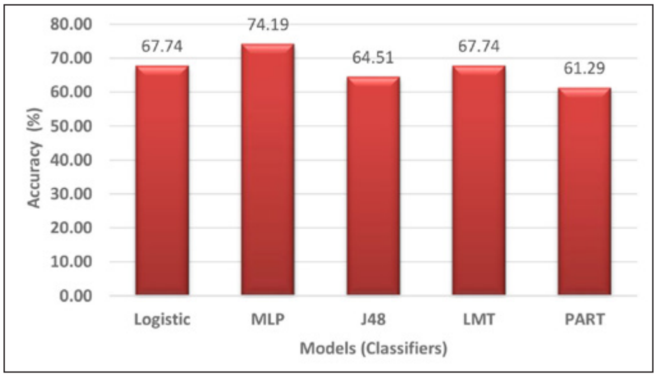


Fig. 3: Correctly classified instances (prediction accuracy) of classification models

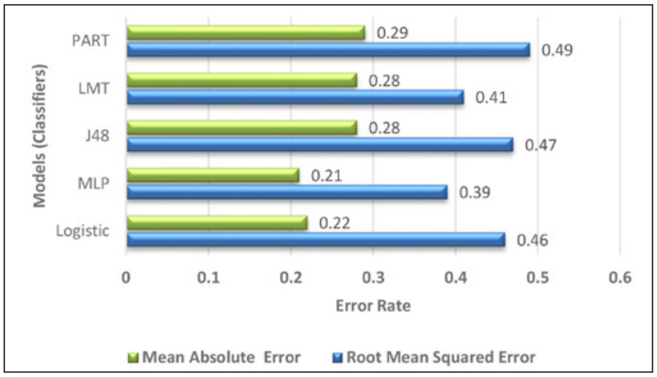


Fig. 4: Error results of classification models

nificant attributes (Z_{538}, Z_{439}, Z_{646}, Z_{132}, Z_{147}, Z_{247}, Z_{536}, Z_{142}, Z_{440}, $Z_{245\ \mathrm{and}}$ Z_{537}) for classification were sorted out using 'cfsSubsetEval' feature selection algorithm as Arunkumar and Ramakrishnan (2017) reported that attribute selection improves the quality of the data and increases the accuracy of data mining algorithms. The results presented in Table 2 indicated that the function based and tree based models have better performance than rule based model. In case of function based models, two models are examined *viz.,* Logistic and MLP**.** MLP has better performance than logistic model. For tree based models, two models are examined *viz*., J48 and LMT. The LMT model has better performance than J48. In general, it could be observed that, MLP model is better than LMT model further underlining that MLP is best fitted model to classify the experimental dataset.

The Fig. 3 shows the prediction accuracy for different classification models. Out of five models used here, MLP model has better rice yield predictability (74.19 %), than other classification models *viz*., Logistic and LMT models with 67.74 % predictability. PART classification recorded lowest predictability (61.29 %). The results were corroborated by Bhojani and Bhatt (2018) also while applying data mining technique in wheat crop yield forecasting and reported that the performance of the MLP model was quite better than the other models used.

MLP model recorded minimal Mean Absolute Error (MAE) of 0.21 and Root Mean Squared Error (RMSE) of 0.39 (Fig.4) during the prediction processes as compared with other models in contrast, to the highest error rate of 0.29 and 0.49 of MAE and RMSE, respectively as exhibited by PART classification. Muradkhanli (2018) developed MLP neural network model using the sigmoid activation function for forecasting the oil production and evaluated the results on the basis of RMSE. MLP model provided better results compared to the regression analysis and concluded that it would be beneficial to Oil Company of Azerbaijan for forecasting the oil production.

The Fig. 5 explains the true positive rate for different classification models. Out of five models used here, MLP model has better true positive rate (0.742) than other classification models followed by Logistic and LMT with 0.677. PART classification has the lowest true positive rate (0.613).

The false positive rate for different classification

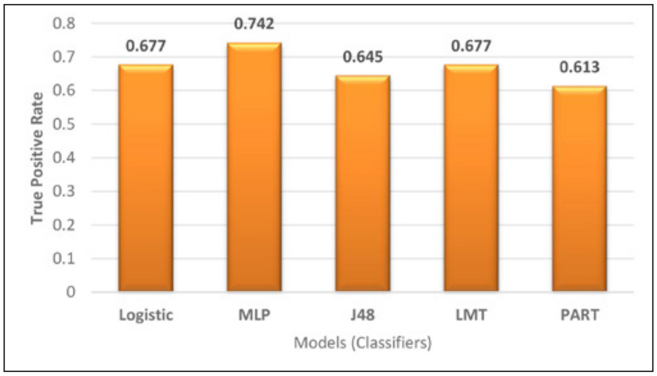


Fig. 5: True positive rate (sensitivity) of classification models

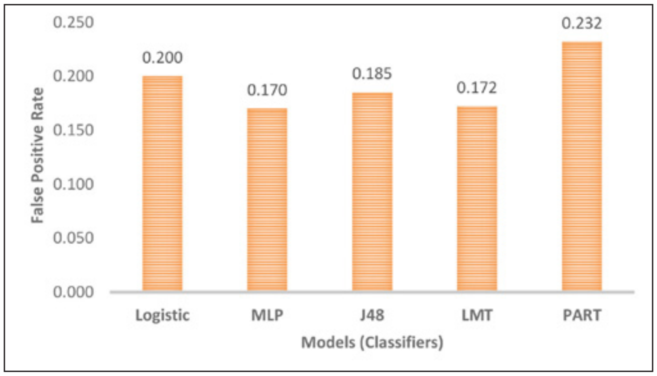


Fig. 6: False positive rate (specificity) of classification models

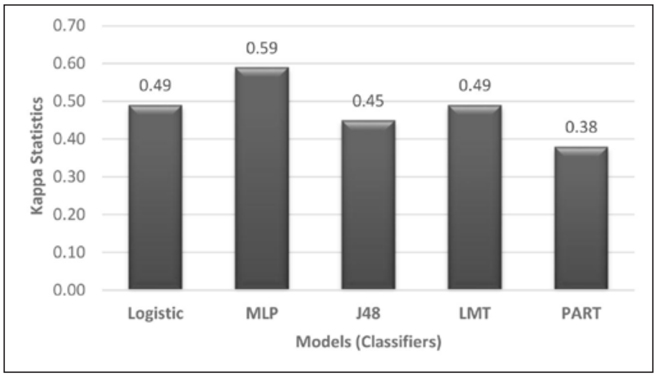


Fig. 7: Kappa statistics of classification models

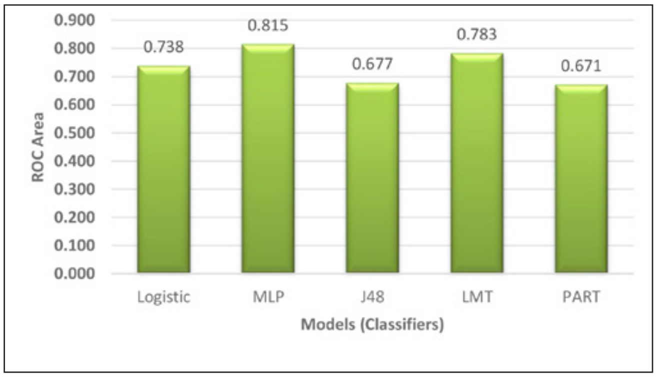


Fig. 8: ROC area of classification models

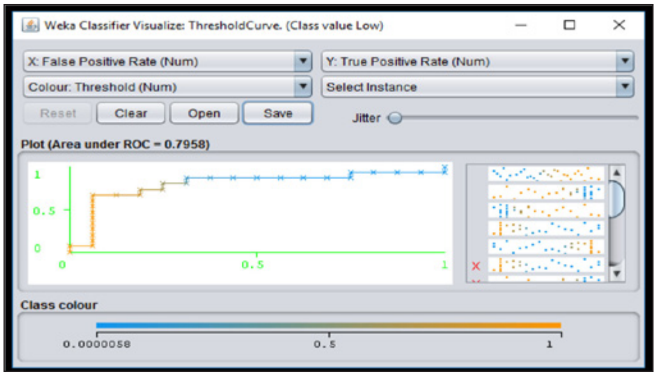
models exhibited similar trend as observed in case of true positive rate. Out of five models used, MLP model recorded lowest false positive rate of 0.170 followed by LMT with 0.172 (Fig. 6). PART classification has the highest false positive rate with 0.232.

The Fig. 7 depicts the kappa statisticsvfor different classification models. Out of five models used, MLP model has better kappa statistics (0.59) than other classification models followed by Logistic and LMT (0.49) models. While, PART classification has the lowest kappa statistics (0.38).

The Fig. 8 explains the Receiver Operating Characteristic (ROC) Area for different classification models. MLP model recorded a highest ROC Area 0.815) followed by LMT (0.783) in contrast of PART with lowest ROC (0.671).

ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: FP Rate (Specificity) on the X axis and TP Rate (Sensitivity) on the Y axis. Fig. 9, 10 and 11 exhibit Area under ROC for class value Low (0.7958), class value Normal (0.7222) and class value High (0.9762) respectively, which found to be good for MLP model only. Hence, it is evident that MPL model performance is better than the other models while predicting rice yields.

The Table 3 explains the statistics after applying the five classification models on rice yield dataset. The fitted MLP model has achieved the highest prediction accuracy of 74.19 %, sensitivity of 0.742 and precision of 0.743. In contrast, PART exhibited the lowest prediction accuracy of 61.29 %, sensitivity of 0.613 and precision of 0.579. In case of specificity, the MLP model has obtained lowest specificity of 0.170 and PART model has obtained highest specificity of 0.232. The F1 score and Mathews Correlation Coefficient were computed to measure the test's accuracy and quality of classification, respectively. The MLP model has achieved the highest F1 score of 0.742 and MCC of 0.581. While, PART has achieved the lowest F1 score of 0.591 and MCC of 0.378. Gandhi *et al*., (2016) too demonstrated the prediction of rice crop yield using Support Vector Machine (SVM), a data mining technique and further showed that the other classification models *viz.,* Naïve Bayes and MLP performed better by achieving the highest accuracy, specificity and sensitivity as compared with SMO (Sequential Minimal Optimization) classification model.



Weka Classifier Visualize: ThresholdCurve. (Class value Low)

X: False Positive Rate (Num)

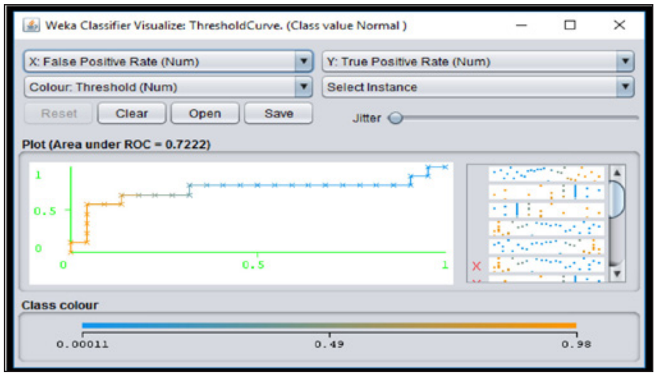
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.7958)

Fig. 9: Thresholds curve (class value low) for MLP



Weka Classifier Visualize: ThresholdCurve. (Class value Normal)

X: False Positive Rate (Num)

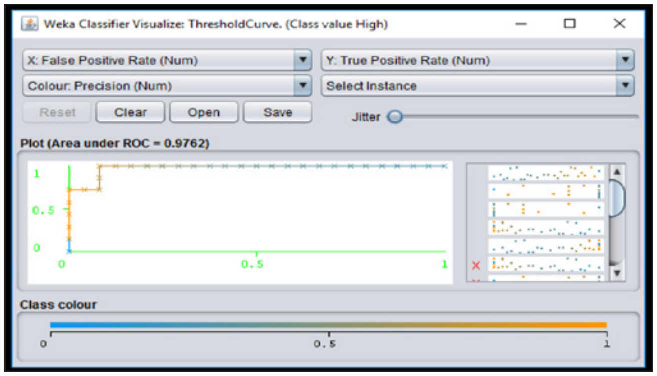
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.7222)

Fig. 10: Thresholds curve (class value normal) for MLP



Weka Classifier Visualize: ThresholdCurve. (Class value High)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Precision (Num)

Select Instance

Plot (Area under ROC = 0.9762)

Fig. 11: Thresholds curve (class value high) for MLP

CONCLUSION

Based on all the benchmarks used to measure the predictability of fitted models employed for rice yield estimation, it was discovered that MLP classifier model performed better by achieving the highest prediction accuracy 74.19 %, sensitivity of 0.742 and precision of 0.743 as compared with other fitted models. The MLP model has achieved the highest F1 score (0.742) and MCC (0.581).Thus, the MLP model explained 74 per cent of total variation in rice crop yield. Hence, it can be concluded that the study helps the researcher in efficient algorithm selection for the rice yield prediction. The outcome of this research would help the policy makers in decision making about the import and export of rice in advance, thus reducing the chaos in its availability for consumption besides strengthening the public distribution system.

REFERENCES

Arunkumar, C. and Ramakrishnan, S. (2017). Performance analysis of classifiers on filter-based feature selection approaches on microarray data. *Bio-Insp. Comput. Inf. Retri. App.*, 41-70.

Biswas, B. and Singh J. (2020). Assessing yield - weather relationships in *kharif* maize under Punjab conditions using data mining method. *J. Agrometeorol.,* 22 (Special Issue): 104 -111.

Bhojani, S.H. and Bhatt, N. (2018). Application of data mining technique for wheat crop yield forecasting for districts of Gujarat state. *Int. J. Sci. Res. Pub.,* 8(7): 302-306.

Diriba, Z. and Borena, B.(2013). Application of data mining techniques for crop productivity prediction. HiLCoE *J. Comp. Sci. Tech*., 1: 151-155.

Fathima, N G and Geetha, R. (2014). Agriculture crop pattern using data mining techniques. *Intl.J. Adv. Res. Comp Sci. Soft. Eng.,* 4: 781-785.

Gandhi, N., Petkar, O. and Armstrong, L. J. (2016). Predicting rice crop yield using bayesian networks. *In*: Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, p.795-799.

Liao, S.H. (2003). Knowledge management technologies and applications literature review from 1995 to 2002. *Exp. Syst.Applicat.,* 25(2):155-164.

Jain M. (2009). Data mining: typical data mining process for predictive modeling. BPB Publications, New Delhi., p.235-241.

Mucherino, A., Papajorgji, P. and Pardolas, P.M. (2009). A survey of data mining techniques applied to agriculture. *Operat. Res: An Intl. J.,* 9 (2):121-140.

Muradkhanli, L. (2018). Neural networks for prediction of oil production. *IFAC Pap. Onl.,* 51(30):415–417.

Rani, A. and Vidyavathi, B.M. (2013). Ameliorated methodology for the design of sugarcane yield prediction using decision tree. Compusoft - *An Int. J. Adv. Comp. Tech.,* 4: 1882-1889.

Singh, P., Kumar, S. and Dadhich, S. (2017). Impact of climate variability on the production and productivity of maize (*Zea mays*) in north western Himalayan region, India. *J. Agrometeorol.,* 19: (Special Issue): 338-44

Sujatha, R. and Isakki, P. (2016). A study on crop yield forecasting using classification techniques. *In:* Proceeding of the International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE), Kovilpatti, India,p.1-4.

Received : September 2020 : Accepted : May 2021