International Journal of Knowledge Based Computer Systems
9 (1) June 2021, 28-32
http://www.publishingindia.com/ijkbcs/

Comparative Analysis of Supervised Machine Learning Techniques in Crop Yield Prediction

Sadakshya Sharma[1*], Jeevan Singh Bhasin[2], Ashish[3], Haneet Kour[4] and Karun Handa[5]
[1]B.Tech. Final Year Student, Department of Computer Science and Engineering, PIET, Haryana, India.
Email: sadakshya.27s@gmail.com
[2]B.Tech. Final Year Student, Department of Computer Science and Engineering, PIET, Haryana, India.
Email: jeevans30121998@gmail.com
[3]B.Tech. Final Year Student, Department of Computer Science and Engineering, PIET, Haryana, India.
Email: ashishghn282@gmail.com
[4]Ph.D. Research Scholar, Department of Computer Science and IT, University of Jammu, J&K, India.
Email: haneetkour9@gmail.com
[5]Assistant Professor, Department of Computer Science and Engineering, PIET, Haryana, India.
Email: karun.cse@piet.co.in
*Corresponding Author

Abstract: Machine learning techniques play an important role in solving real world problems. These techniques are also found to be successful in the field of Agriculture for crop yield prediction, leaf disease detection, fruit disease detection, vegetable quality assessment, etc. In this paper, the authors performed comparative analysis of various supervised machine learning techniques for crop yield prediction from soil parameters. Five supervised machine learning techniques such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT) and Random Forest (RF) have been taken for the experimental analysis. The experiments have been carried out for the prediction of the three most commonly grown crops in India: Rice, Wheat and Mustard. The performance of each technique for every crop taken in this study, has been evaluated on the basis of four metrics i.e. accuracy, recall, precision and f-score. The experimental results revealed that decision tree and random forest performed better than all the other supervised machine learning techniques taken in this study, for the prediction of each crop.

Keywords: DT, KNN, Mustard, RF, Rice, SVM, Wheat.

I. Introduction

Agriculture sector plays an important role in the economy of any country. In India, agriculture comprises over 60% of the Indian population but it contributes only about 18% of the total GDP. The imbalance between the dependence and contribution is due to the lack of proper knowledge and implementation of advanced techniques in agriculture by farmers at every step of crop production [1]. A significant activity for decision-makers is the estimation of crop yields. A precise model for prediction of crop yield can enable farmers to make a decision on when to cultivate and when to grow. In recent years, Machine Learning (ML) has emerged as a useful tool for forecasting crop yields, as well as recommending which crops to grow (Klompenburg *et al.*, 2020) [5]. The current research work makes use of supervised machine learning algorithms to predict the yield of most commonly grown crops i.e. Rice, Wheat and Mustard. In this study, the authors performed comparative analysis of various supervised machine learning techniques such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT) and Random Forest (RF) for crop yield prediction. The main aim of the current study is to reveal out which machine learning technique is better for yield prediction of rice, wheat and mustard crops.

This research paper is distributed as: Section II represents previous work related with current study, Section III explains materials and methodology adopted for the experimental set-up, Section IV shows experimental results and analysis, and Section V describes conclusion and future work.

II. Literature Review

In recent past, significant research has been done in the domain of agriculture using machine learning techniques to predict production of different crops. Bhuyar (2014) [2] applied J48, Naïve Bayes and random forest techniques for the prediction

of soil fertility. The experimental results predicted that J48 algorithm performed better with an accuracy of 98.17% than other algorithms taken in the study.

A system for soil classification based on Naive Bayes and K-Nearest Neighbor supervised learning techniques, has been proposed by Supriya (2017) [12]. In the same year, Shastry *et al.* (2017) [10] applied various regression techniques such as quadratic, pure quadratic, interaction and polynomial for crop yield prediction. The authors have taken wheat, maize and cotton crops for experimental study. Maize dataset had 78 records with 04 attributes, wheat dataset had 50 records with 05 attributes; and cotton dataset had 123 records with 08 attributes. Among the entire regression models taken in the study, polynomial model predicted better results for wheat crop, pure-quadratic model predicted better results for maize crop, and the proposed SLR model predicted better results for cotton crop. A crop yield prediction model based on association rule mining was implemented by Manjula and Djodiltachoumy (2017) [8]. The experiments have been conducted on dataset of Tamil Nadu state. The dataset had 08 input attributes and 01 output attribute. The proposed approach predicted accuracy of approximate 85%.

Next year, Priya *et al.* (2018) [9] implemented random forest technique for the yield prediction of rice crop. The experiment was conducted on the rice dataset with 05 attributes such as rainfall, maximum temperature, crop production, etc. The dataset was divided into training set and testing set in the ratio of 67:33. In the same year, Neupper rule based algorithm was proposed by Manimekalai and Nandhini (2018) [7] for crop yield prediction from soil parameters. The proposed approach was implemented by combining ANN with ripple classifier. This method worked by generating weights of ANN through input parameters of the dataset; followed by construction of decision tree based on the weight values of soil parameters through Ripper classifier procedure. The Ripper procedure has two steps: build rule and prune rule. The proposed model predicted accuracy of 96.40% with precision of 95% and f-score of 0.9599.

Lata and Chaudhari (2019) [6] developed decision support system for crop yield prediction based on data mining techniques. The datasets for kharif and rabi crops have been taken for experimental study. The proposed model worked by integrating feature selection and classification technique for crop yield prediction. Bondre and Mahagaonkar (2019) [3] applied support vector machine and random forest for crop yield prediction from soil parameters. The authors also integrated fertilizer recommendation in case of low crop yield. The experimental results predicted average accuracy of 97.48% for random forest and 99.47% for support vector classifier.

Champaneri *et al.* (2020) [4] implemented a model based on random forest technique for the yield prediction of crop. The authors collected crop dataset of Maharashtra state. The experiment was conducted on the collected dataset with climatic attributes such as precipitation, temperature, cloud cover, vapor pressure, and wet day frequency. The dataset was divided into training set and testing set in the ratio of 75:25. In the same year, Suganya *et al.* (2020) [11] performed comparative study of various supervised machine learning techniques for crop yield study. The authors have taken dataset of various crops for experimental study. The experimental results revealed that logistic regression predicted better results among all the techniques taken in the study.

III. Materials and Methodology

The main objective of the current research is to perform comparative analysis of different supervised machine learning techniques for crop yield prediction. To attain the objective of the present study, experiments have been conducted on Matlab platform. For the current study, three most commonly grown crops i.e. Rice, Wheat and Mustard have been undertaken for experimental analysis. The data for each crop has been taken from "*www.soilhealth.dac.gov.in*" under Model Village Programme 2019-20. The data set for each crop consists of 11 input attributes and one output attribute. The input attributes of the dataset representing soil nutrients status and these attributes are *Ph* (ph value of soil), *EC* (electrical conductivity), *OC* (organic carbon), *N* (nitrogen), *P* (phosphorus), *K* (potassium), *S* (sulphur), *Cu* (copper), *Fe* (iron), *Zn* (zinc) and *Mn* (manganese). The output attribute represents three classes namely low, medium and high for each crop yield. The dataset comprises 5749 records for wheat crop, 5753 records for rice crop and 6134 records for mustard crop. The details of the dataset for each crop have been presented in Table I. After data collection, pre-processing has been performed to transform nominal values into numerical form, to impute missing data using *mean* statistics, and to normalize the data in the range [0, 1] using *min-max normalization*. The overall methodology for the present work is presented in Fig. 1.

Table I: Details of Dataset Output Classes for Each Crop

Crop	Output Class Label			Total Records
	Low	Medium	High	
Wheat	3947	1519	283	5749
Rice	2952	2058	743	5753
Mustard	4743	685	706	6134

The experiments have been performed in three stages. The first stage deals with implementation of undertaken ML techniques for yield prediction of Wheat crop, the second stage deals with Rice crop yield prediction, and the third stage deals with implementation of Mustard crop. For each crop, the data set has been divided into *Train Set* and *Test Set* in the ratio of 80:20. For *wheat crop*, train set and test set have 4599 instances and 1150 instances respectively. In case of *rice crop*, train set and test set have 4602 instances and 1151 instances respectively. In case of *mustard crop*, train set and

test set have 4907 instances and 1227 instances respectively. Five supervised machine learning techniques such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT) and Random Forest (RF) have been implemented for yield prediction of undertaken crops in order to train the model on train set. Each model has been trained with 10-fold cross validation. KNN model has been implemented with *minkowski* distance measure and 5 neighbors. Each trained model has been validated on test set to evaluate the model performance. The performance of all the trained models has been assessed using four measures such as *accuracy*, *recall*, *precision*, and *F-score*. On the basis of the experimental results, comparative analysis of all the trained models has been carried out to reveal the most accurate technique for each undertaken crop yield prediction.

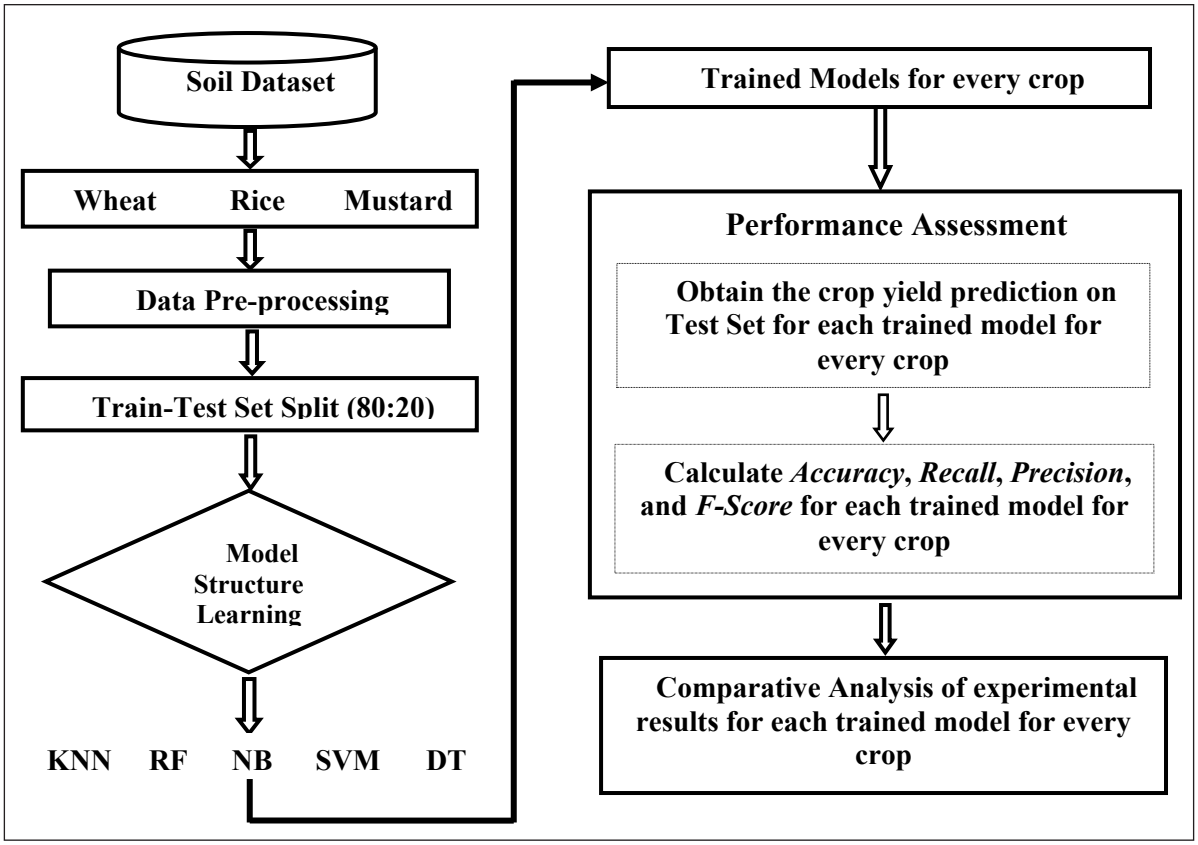

Fig. 1: Flowchart for the Current Study

IV. Results and Discussion

From experimental results, it can be found out that all ML techniques under study can be used for crop yield prediction. Table II presents the experimental results for all supervised machine learning techniques taken in the study for wheat crop yield prediction. The performance of all the techniques has been evaluated on the basis of four parameters i.e. *accuracy*, *recall*, *precision* and *f-score*. Among all the ML techniques under study, Naïve Bayes achieved lowest accuracy of 80% with f-score of 0.71, whereas *Decision Tree* and *Random Forest* predicted highest accuracy of 99% with f-score of 0.98 and 0.99 respectively. Fig. 2 presents the comparative analysis of ML techniques for *wheat crop* yield prediction. From Fig. 2, it has been observed that Naïve Bayes predicted lowest performance for wheat yield prediction as it achieved lowest accuracy, recall, precision and f-score. *Decision Tree* and *Random Forest* have been found to be best approaches for wheat yield prediction as it achieved highest accuracy, recall, precision and f-score.

Table II: Experimental Results for Wheat Crop

Classifier	Accuracy	Precision	Recall	F-Score
KNN	93%	91%	94%	0.92
SVM	89%	85%	90%	0.87
NB	80%	67%	80%	0.71
DT	99%	98%	97%	0.98
RF	99%	99%	99%	0.99

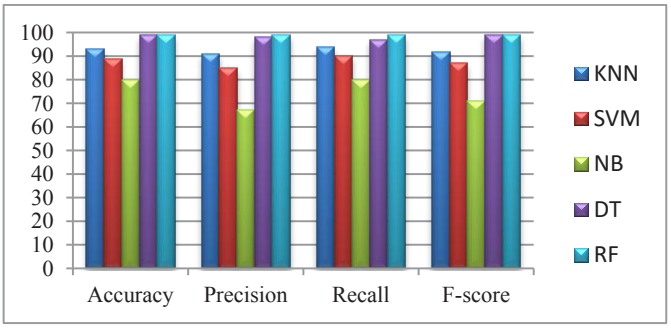

Fig. 2: Comparison of ML Techniques under Study for Wheat Crop

Table III presents the experimental results for all supervised machine learning techniques taken in the study for *rice crop* yield prediction. Among all the ML techniques under study, Naïve Bayes, SVM and KNN achieved less accuracy of 78%, 78% and 79% with f-score of 0.77, 0.79 and 0.80 respectively. *Decision Tree* and *Random Forest* predicted highest accuracy of 99% with f-score of 0.99. Fig. 3 presents the comparative analysis of ML techniques for rice crop yield prediction. From Fig. 3, it has been observed that both Naïve Bayes and SVM achieved accuracy of 78%, and *the former* achieved low precision and high recall with f-score of 0.77 whereas *the latter* achieved high precision and low recall with f-score of 0.79. Thus SVM has been found to be slightly better than Naïve Bayes for rice yield prediction. *Decision Tree* and *Random Forest* have been found to be best approaches for rice yield prediction as it achieved highest accuracy, recall, precision and f-score.

Table III: Experimental Results for Rice Crop

Classifier	Accuracy	Precision	Recall	F1-Score
KNN	79%	79%	81%	0.80
SVM	78%	80%	79%	0.79
NB	78%	75%	80%	0.77
DT	99%	99%	99%	0.99
RF	99%	99%	99%	0.99

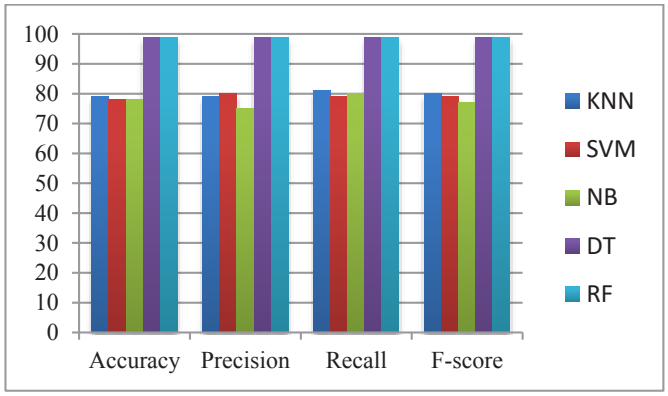

Fig. 3: Comparison of ML Techniques under Study for Rice Crop

Table IV presents the experimental results for all supervised machine learning techniques taken in the study for *mustard crop* yield prediction. Among all the ML techniques under study, Naïve Bayes achieved least accuracy of 53% with f-score of 0.54. *Decision Tree* and *Random Forest* again predicted highest accuracy of 99% with f-score of 0.98. Fig. 4 presents the comparative analysis of ML techniques for mustard crop yield prediction. From Fig. 4, it has been observed that both Naïve Bayes and SVM achieved accuracy of 53% and 85%, and *the former* achieved low precision and high recall with f-score of 0.54 whereas *the latter* achieved high precision and low recall with f-score of 0.64. Thus SVM has been found to be better than Naïve Bayes for rice yield prediction. *Decision Tree* and *Random Forest* have been found to be best approaches for rice yield prediction as it achieved highest accuracy, recall, precision and f-score.

Table IV: Experimental Results for Mustard Crop

Classifier	Accuracy	Precision	Recall	F1-Score
KNN	88%	81%	76%	0.78
SVM	85%	80%	62%	0.64
NB	53%	58%	70%	0.54
DT	99%	98%	99%	0.98
RF	99%	98%	98%	0.98

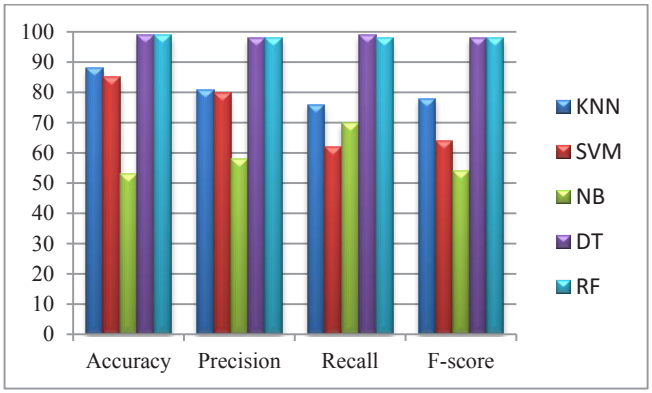

Fig. 4: Comparison of ML Techniques under Study for Mustard Crop

From Tables II, III and IV, it has been analyzed that both *Random Forest* and *Decision Tree* have predicted better performance as compared to the other ML techniques taken in the study for each crop. Thus these two ML techniques are very effective for crop yield prediction. Naïve Bayes technique predicted lowest performance for every crop taken in this study. Thus this technique is least effective for crop yield prediction.

V. Conclusion and Future Scope

From the experimental study, it is concluded that machine learning techniques can be efficiently used for crop yield prediction. But, in this study, DT and RF have been found to

be the most precise procedures for crop yield prediction. These effective ML techniques will help the farmers in predicting yield in advance based on soil parameters.

References

[1] Agriculture in India - Statistics & Facts. Accessed: Nov. 30, 2020. [Online]. Available: https://www.statista.com/topics/4868/agricultural-sector-in-india/

[2] V. Bhuyar, "Comparative analysis of classification techniques on soil data to predict fertility rate for Aurangabad district," *International Journal of Emerging Trends and Technology in Computer Science*, vol. 3, no. 2, pp. 200-203, 2014.

[3] D. A. Bondre, and S. Mahagaonkar, "Prediction of crop yield and fertilizer recommendation using machine learning algorithms," *International Journal of Engineering Applied Sciences and Technology*, vol. 4, no. 5, pp. 371-376, 2019.

[4] M. Champaneri, C. Chandvidkar, D. Chachpara, and M. Rathod, "Crop yield prediction using machine learning," *International Journal of Science and Research*, vol. 9, no. 4, pp. 645-648, 2020.

[5] T. V. Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, 2020, Art. no. 105709.

[6] K. Lata, and B. Chaudhari, "Crop yield prediction using data mining techniques and machine learning models for decision support system," *Journal of Emerging Technologies and Innovative Research*, vol. 6, no. 4, pp. 391-396, 2019.

[7] S. Manimekalai, and K. Nandhini, "Crop yield prediction from soil parameters through Neupper rule established algorithm," *International Journal of Engineering and Technology*, vol. 7, no. 3.34, pp. 908-912, 2018.

[8] E. Manjula, and S. Djodiltachoumy, "A model for prediction of crop yield," *International Journal of Computational Intelligence and Informatics*, vol. 6, no. 4, pp. 298-305, 2017.

[9] P. Priya, U. Muthaiah, and M. Balamurugan, "Predicting yield of the crop using machine learning algorithm," *International Journal of Engineering Sciences and Research Technology*, vol. 7, no. 4, pp. 1-7, 2018.

[10] A. Shastry, H. A. Sanjay, and E. Bhanusree, "Prediction of crop yield using regression techniques," *International Journal of Soft Computing*, vol. 12, no. 2, pp. 96-102, 2017.

[11] M. Suganya, R. Dayana, and R. Revathi, "Crop yield prediction using supervised learning techniques," *International Journal of Computer Engineering & Technology*, vol. 11, no. 2, pp. 9-20, 2020.

[12] D. M. Supriya, "Analysis of soil behavior and prediction of crop yield using data mining approach," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, pp. 9648-9652, 2017.