Volume 64, Issue 2, 2020
Journal of Scientific Research
Institute of Science,
Banaras Hindu University, Varanasi, India.

# Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis

Vaishali Pandith[*1], Haneet Kour[1], Surjeet Singh[2], Jatinder Manhas[3], and Vinod Sharma[1]

[1]Department of Computer Science and IT, University of Jammu.
vaishalipandith@gmail.com*, haneetkour9@gmail.com, vnodshrma@gmail.com
[2]Department of Computer Science, G.M.N. (P.G.) College, Ambala Cantt. surjeetsagwal@gmail.com
[3]Department of Computer Science and IT, Bhaderwah Campus, University of Jammu. manhas.jatinder@gmail.com

***Abstract:*** **Soil is an important parameter affecting crop yield prediction. Analysis of soil nutrients can aid farmers and soil analysts to get higher yield of the crops by making prior arrangements. In this paper, various machine learning techniques have been implemented in order to predict Mustard Crop yield in advance from soil analysis. Data for the experimental set-up has been collected from Department of Agriculture Department, Talab Tillo, Jammu; comprising soil samples of different districts of Jammu region for Mustard crop. For the current study, five supervised machine learning techniques namely K-Nearest Neighbor (KNN), Naïve Bayes, Multinomial Logistic Regression, Artificial Neural Network (ANN) and Random Forest have been applied on the collected data. To assess the performance of each technique under study; five parameters namely accuracy, recall, precision, specificity and f-score have been evaluated. Experimentation has been carried out to make known the most accurate technique for mustard crop yield prediction. From experimental results, it has been predicted that KNN and ANN (among the undertaken ML techniques for the study) found to be most accurate techniques for mustard crop yield prediction.**

***Index Terms:*** **ANN, KNN, Machine Learning, Mustard Crop, Naïve Bayes, Random Forest.**

## I. INTRODUCTION

Machine Learning is a technology that provides systems the ability to automatically learn and improve from experience by repeatedly training. It includes set of well-defined models that collect specific data and apply specific algorithms to achieve desired results. Machine learning techniques have been applied to agriculture domain in order to improve the productivity and quality of the crops grown. The algorithms in Machine Learning are used to determine for a particular crop under which conditions the best yield would be produced.

Crop Yield prediction depends on various factors like the soil composition, type, climate, region geography and disease or pests. Soil is a very important factor affecting plant growth. It consists of solids (minerals and organic matter), liquids (water and solutes) and gases (mainly oxygen and carbon dioxide) and contains living organisms. All these elements provide their physical and chemical properties. To maintain fertility, to achieve better yield, and to protect the environment; it is necessary to nurture the soil properly. The analysis of soil nutrients is very useful for the farmers in determining the type of yield to be grown in a particular soil condition. Good soil fertility management requires careful identification of the limits of current nutritional deficiencies and monitoring of changes in soil fertility to predict their shortage. These gaps must be mitigated by sound and best practices in terms of nutrients, water, plants and energy for soil management, in order to maintain food production at a reasonable level to ensure high productivity at the same time. Therefore, managing soil fertility at optimal levels is one of the key factors for achieving high and sustainable productivity (Smriti 2015).

*Rapeseed-Mustard* is the second most important oil seed crop. Productivity of rapeseed mustard in J&K remains unstable from past few years (Rakesh 2018). The current research work makes use of supervised machine learning algorithms to predict mustard crop yield for different districts of Jammu region. In this work, the authors implemented different ML techniques for crop yield prediction to find out the most accurate techniques for crop yield prediction under study.

This research paper is organized as follows: Section 2 represents Literature Review, Section 3 explains Materials and Methodology, Section 4 shows Results and Discussion, and Section 5 describes Conclusion and Future Scope.

* Corresponding Author

## II. LITERATURE REVIEW

In recent years, machine learning techniques have been applied in agriculture domain by various researchers. A review on the implementation of different ML techniques in the field of crop yield prediction from soil analysis from past few years is presented as under.

*Bhuyar* (2014) proposed an approach where different classification algorithms such as J48, Naïve Bayes, and Random forest algorithm were applied to soil data set to predict its fertility. J48 algorithm gave better result with an accuracy of 98.17% than other algorithms.

After two years, *Rajeshwari and Arunesh* (2016) performed a comparative analysis of ML algorithms i.e. Naive Bayes, JRIP and J48 for prediction of soil types. The experiments were performed on soil data consisting of 110 samples using data analytics tool R. The experimental results predicted that JRIP algorithm performed better as it gave highest accuracy of 98.18% with kappa statistic of approximate 1.0.

In the same year, *Sujata* (2016) proposed a model to estimate the crop yield in order to improve the value and gain of farming area using data mining techniques.

*Awasthi and Bansal* (2017) performed comparative study on two data mining techniques namely Artificial Neural Network and Support Vector Machine with the help of data analytics tool R. ANN was implemented with 7 hidden nodes and this model trained for 73073 steps. It predicted accuracy of 55% with root mean square error is of 15. SVM implemented with Radial basis kernel and it achieved much better results with 74% of accuracy.

In the same year, *Singh et al.* (2017) implemented different machine learning techniques in order to predict the category of the rice crop yield based on macro-nutrients and micro- nutrients status in dataset. ANN, Naïve Bayes and KNN are applied on soil data. Decision Tree Classifier and Naïve Bayes Classifier are found to be better models for classifying the soils into categories and in the prediction of yield on the basis of Nutrient status in the soil.

*Supriya* (2017) presented a system based on data mining techniques in order to predict the category of the analyzed soil datasets (yield of crop). Naive Bayes and K-Nearest Neighbor methods are used. System architecture is also developed and tested using data mining technology.

*Verma et al.* (2018) also proposed an approach for wheat crop prediction using fuzzy-c means clustering and neural network.

In the same year, *Priya et al.* (2018) implemented Random Forest technique for crop yield prediction in Tamil Nadu state. The dataset for the study includes various parameters such as rainfall, temperature, crop production, etc. and experiments were performed using R Studio.

Next year, *Renuka and Terdal* (2019) applied machine learning techniques namely KNN, SVM and Decision Tree for yield prediction of sugarcane crop. The study was carried out in Python platform. Decision Tree predicted highest accuracy of 99% with less mean square error.

*Jayalakshmi and Devi* (2019) applied ML techniques for predicting soil fertility. The authors implemented C5.0, Random forest and K-Nearest Neighbor for crop production with high accuracy and efficiency by generating a model which predicts whether soil is "Ideal" or "Not Ideal" for crop production based on soil parameter. C5.0 predicted highest performance with an accuracy of 96%.

## III. MATERIALS AND METHODOLOGY

The main objective of the current research work is to predict *mustard crop* yield from soil analysis using machine learning techniques. To attain the objective of the current research, experiments have been carried out on Matlab platform. After gaining insight of problem domain, discussion with farmers and soil chemists and reviewing literature; research problem has been framed out. For current research problem, real data has been collected from *Soil Testing Lab, Directorate of Agriculture Department, Talab Tillo, Jammu*. Data has been collected from mustard growing areas of various districts of Jammu Region under *Model Village Programme* 2019-20. This dataset consists of 5000 instances with 11 input parameters representing soil nutrient status of Jammu region and one output attribute (i.e. Class Label). The parameters of the dataset 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) representing soil nutrients. The output attribute represents three classes for mustard crop yield namely low, medium and high. Out of total 5000 instances collected, 3666 falls in *Low* class, 958 falls in class *Medium* and 376 falls in class *High*. The first 15 instances of the dataset are presented in table I.

Table I. First 15 records of the Dataset

| pH | EC | OC | N | P | K | S | Cu | Fe | Zn | Mn | O/P |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 6.9 | 0.53 | 0.52 | 361.4 | 13.3 | 242.7 | 13.4 | 0.68 | 6.4 | 0.36 | 2.72 | 1 |
| 7.3 | 0.51 | 0.43 | 298.8 | 16.6 | 258.4 | 12.6 | 0.60 | 6.4 | 0.34 | 2.84 | 1 |
| 7.3 | 0.63 | 0.23 | 166.8 | 12.3 | 252.2 | 16.2 | 0.62 | 6.6 | 0.36 | 2.76 | 0 |
| 7.2 | 0.52 | 0.36 | 250.2 | 16.3 | 264.3 | 11.8 | 0.74 | 6.3 | 0.32 | 2.64 | 0 |
| 7.2 | 0.49 | 0.44 | 305.8 | 12.5 | 228.8 | 19.4 | 0.68 | 7.5 | 0.54 | 2.72 | 1 |
| 7.1 | 0.73 | 0.22 | 152.9 | 14.7 | 252.2 | 17.8 | 0.74 | 7.1 | 0.32 | 2.86 | 0 |
| 6.8 | 0.75 | 0.34 | 236.3 | 16.6 | 258.4 | 11.2 | 0.76 | 7.1 | 0.44 | 2.58 | 0 |
| 6.8 | 0.69 | 0.26 | 180.7 | 19.4 | 264.1 | 10.4 | 0.68 | 7.5 | 0.36 | 2.36 | 0 |
| 6.9 | 0.71 | 0.43 | 295.3 | 17.8 | 242.7 | 17.7 | 0.68 | 7.5 | 0.36 | 2.36 | 1 |
| 7.3 | 0.68 | 0.24 | 166.8 | 12.5 | 305.8 | 12.8 | 0.64 | 6.5 | 0.58 | 2.24 | 0 |
| 7.3 | 0.62 | 0.32 | 222.4 | 14.7 | 276.4 | 17.1 | 0.70 | 6.8 | 0.56 | 2.28 | 0 |
| 7.3 | 0.69 | 0.66 | 313.5 | 64.1 | 231.8 | 10.2 | 3.3 | 2.6 | 0.88 | 2.66 | 1 |
| 7.3 | 0.66 | 0.52 | 361.4 | 13.3 | 258.2 | 13.3 | 0.72 | 6.6 | 0.62 | 2.42 | 1 |
| 7.4 | 0.66 | 0.28 | 194.6 | 16.8 | 246.3 | 16.6 | 0.74 | 6.6 | 0.44 | 2.56 | 0 |
| 7.2 | 0.53 | 0.42 | 291.9 | 14.6 | 228.8 | 10.5 | 0.68 | 6.7 | 0.68 | 2.76 | 1 |

After data collection, pre-processing has been performed to transform nominal values into numerical form, to impute missing data and to detect outliers present in the data using statistical technique. Min-max normalization has been done to transform all the input attributes in the range [0, 1]. The overall

methodology for the proposed work is presented in figure 1.

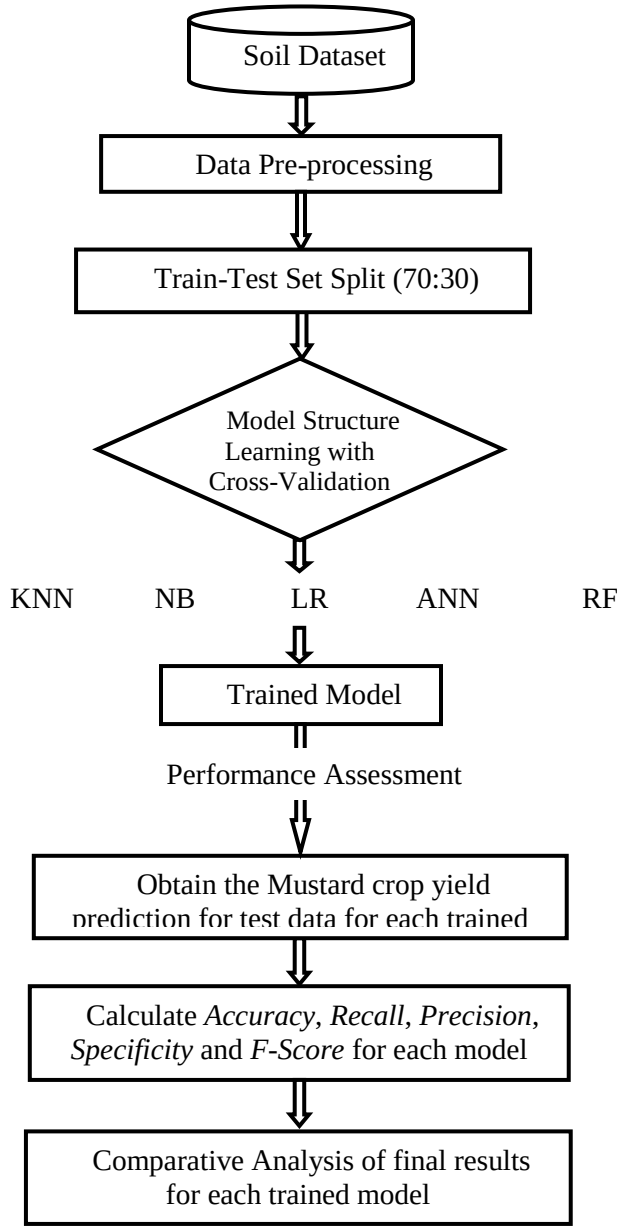


Fig. 1. Flowchart for the current study

The experiments have been carried out in five phases. In the first phase, k-nearest neighbor (KNN) technique has been implemented to train the model for mustard crop yield prediction. The second, third, fourth, and fifth phase deals with implementation of Naive Bayes (NB), Multinomial Logistic Regression (LR), artificial neural network (ANN), and Random Forest (RF) respectively for training mustard crop yield classifier. In each phase, data (with selected features) has been divided into two sets: *training set* and *test set* in the ratio of 70:30; thus training set has 3500 samples and test set 1500 samples. Each trained model has been validated with test data in order to assess its performance. The performance of all the trained models has been measured using performance measures namely *accuracy*, *recall*, *precision*, *specificity* 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 mustard crop yield prediction.

### *A. K-Nearest Neighbor Implementation*

KNN is used for both classification and regression problems. It is one of the simplest classification algorithms. It works by determination of the parameter ***k*** which is number of nearest neighbors. When there is new data point to classify, then its k-nearest neighbors is find out from the training data by calculating the distance between the input variable and the all the data points in the dataset. This distance is calculated using various measures such as *Euclidean distance*, *Minkowski distance*, *Mahalanobis distance*. The larger is k; the better is classification (Harrison 2018, Brownlee 2016).

For the experimental study, KNN has been implemented with 10-fold cross validation and optimum value of k is found to be 25.

### *B. Naïve Bayes Classifier Implementation*

A Naive Bayes classifier is one of the classifiers in a family of simple probabilistic classification techniques in machine learning. It is based on the Bayes theorem with independence features. Each class labels are estimated through probability of given instance. It needs only small amount of training data to predict class label necessary for classification. Naïve Bayes is particularly effective for data sets containing multiclass predictors (Gandhi 2018, Shubam 2018). For the current study, Naïve Bayes has been implemented with 10-fold cross validation.

### *C. Multinomial Logistic Regression Implementation*

It is also called as multiclass classification. Target variable can take more than two value and values should be ordered. The multiclass logistic regression is similar to binary logistic regression, except the label (class) is now an integer in {1, 2, … C} where C is number of classes. Scores for all classes are calculated. It is implemented using *Softmax* function. Class with most votes is chosen for prediction (Martin, Nagesh 2019).

For the experimental study, Logistic regression on all parameters for multiclass labels has been implemented in Matlab using function *mnrfit* and *mnrval* returns the predicted probabilities for the multinomial logistic regression model with predictors X, and the coefficient estimates, B.

### *D. Artificial Neural Network Implementation*

ANN is one of the most used techniques for the prediction model. ANN is usually based on imitation of human brain; just

like our brain it has neurons for transmitting one data to another. All the neurons are connected together in layers. The application of ANN is widely used in agriculture practices. It compares patterns nonlinear effect and underline concept of the relation between them and hence it is a kind of ML technique which has a vast memory. One of the disadvantages of ANN is that where the dataset is significantly different compared to trained data set (Chauhan 2019, Hardesty 2017).

For experimental set up, ANN has been implemented with *Scale Conjugate Gradient algorithm* (trainscg) with 10-fold cross validation for mustard crop yield prediction. ANN model consists of 5 input layer neurons, 21 neurons in hidden layer and 3 output layer neurons.

### *E. Random Forest Implementation*

Random Forest is a supervised learning algorithm. It creates a forest and makes it somehow random. Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks which form the majority of current machine learning systems (Yiu 2019, Donges 2019).

For the current study, *fitrensemble* has been applied on collected dataset with 10-fold cross validation.

## IV. RESULT AND DISCUSSION

From experimental results, it can be found out that all ML techniques under study can be used for crop yield prediction. KNN and random forest predicted *highest accuracy* of 88.67% and 94.13% respectively whereas Naïve Bayes predicted *lowest accuracy* of 72.33%. In terms of *precision*, ANN predicted highest value of 99.94% whereas Logistic regression predicted lowest value of 24.17%. All the classifiers under study except Naïve Bayes, predicted *recall* value more than 90%. It means Naïve Bayes gave highest false negative rate; and Logistic regression gave high false positive rate with lowest true negative rate. ANN and KNN produced higher *specificity* of 99.78% and 80.72% respectively and also highest *f-score* with value of 0.9976 and 0.8405 respectively.

Table II presents the comparative analysis for all ML techniques under study. These techniques have been evaluated on the basis of five parameters namely *accuracy*, *precision*, *recall*, *specificity* and *f-score*. Among all the undertaken ML techniques for the study, KNN and ANN predicted best performance. Figure 2 also presents the summary of comparative results of all ML techniques under study.

Table II. Experimental Results for ML Techniques under Study

| ML Algo | Accuracy | Precision | Recall | Specificity | F-Score |
|---|---|---|---|---|---|
| KNN | 88.67% | 78.14% | 90.92% | 80.72% | 0.8405 |
| NB | 72.33% | 70.91% | 72.69% | 75% | 0.7179 |
| MLR | 80.24% | 24.17% | 96.66% | 0.0000% | 0.3866 |
| ANN | 76.86% | 99.94% | 99.61% | 99.78% | 0.9976 |
| RF | 94.13% | 66.55% | 99.66% | 100% | 0.7981 |

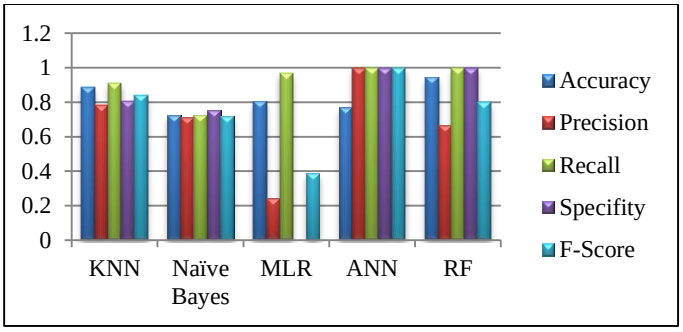


Fig. 2. Comparison of ML techniques under study

## CONCLUSION AND FUTURE SCOPE

From the experimental study, it can be concluded that ML techniques can be effectively used for yield prediction of mustard crop. But, in this study, KNN and ANN are found to be most accurate techniques for mustard crop yield prediction. These effective ML techniques will help the farmers in predicting yield in advance based on soil parameters. In future, Crop yield prediction with huge soil data set can be implemented in Big Data environment. On the basis of results of yield prediction, fertilizer recommendations can also be implemented to help the soil analysts and farmers to take decisions accordingly in case of *low* crop yield prediction.

## ACKNOWLEDGMENT

Authors would like to thank Mr. Ashwani Kotwal, Soil Chemist, Department of Agriculture, Talab Tillo, Jammu for his cooperation for providing data and providing useful guidance for research.

## REFERENCES

Awasthi, N. & Bansal, A. (2017). Application of Data Mining Classification Techniques on Soil Data using R. International Journal of Advances in Electronics and Computer Science, 4, 33-37.

Bhuyar, V. (2014). Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District. International Journal of Emerging Trends & Technology in Computer Science, 3(2), 200-203.

Brownlee, J. (2016). K Nearest Neighbors for Machine Learning. Retrieved March 23, 2020, from https://machinelearningmastery.com.

Chauhan, N. S. (2019). Introduction to Artificial Neural Networks (ANN). Retrieved March 26, 2020, from https://towardsdatascience.com.

Donges, N. (2019). A Complete Guide to the Random Forest Algorithm. Retrieved March 30, 2020, from https://builtin.com/data-science/random-forest-algorithm.html.

Gandhi, R. (2018). Naïve Bayes Classifier. Retrieved March 25, 2020, from https://towardsdatascience.com.

Hardesty. L. (2017). Explained: Neural Networks. Retrieved March 26, 2020, from https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414.

Harrison, O. (2018). Machine Learning Basics with the K-Nearest Neighbors Algorithm. Retrieved March 23, 2020, from https://towardsdatascience.com.

Jayalakshmi, R. & Devi, M. S. (2019). Relevance of Machine Learning Algorithms on Soil Fertility Prediction using R. International Journal of Computational Intelligence and Informatics, 8(4), 193-199.

Martin, K. Logistic Regression Models for Multinomial and Ordinal Variables. Retrieved March 27, 2020, from https://www.theanalysisfactor.com/logistic-regression-models-for -multinomial-and-ordinal-variables.html

Nagesh, S. (2019). Real world implementation of Logistic Regression. Retrieved March 27, 2020, from http://towardsdatascience.com.

Priya, P., Muthaiah, U. & Balamurugan, M. (2018). Predicting Yield of the Crop Using Machine Learning Algorithm. International Journal of Engineering Sciences and Research Technology, 1-7.

Rajeshwari, V. & Arunesh, K. (2016). Analyzing Soil Data using Data Mining Classification Techniques. Indian Journal of Science and Technology, 9, 1-4.

Rakesh (2018). Performance analysis of rapeseed-mustard under different agro-climatic conditions of Jammu division of J&K state.

Renuka & Terdal, S. (2019). Evaluation of Machine Learning Algorithms for Crop Yield Prediction. International Journal of Engineering and Advanced Technology, 8(6), 4082-4086.

Shubham, J. (2018). Naïve Bayes Theorem. Retrieved March 25, 2020, from https://becominghuman.ai/naïve-bayes-theorem-d8854a41ea08.html.

Singh, V., Sarwar, A. & Sharma, V. (2017). Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach**.** International Journal of Advanced Research in Computer Science, 8(5), 1254-1259.

Smriti (2015). A Review on Soil Property Detection using Machine Learning Approach. International Journal Online of Science, 4.

Sujatha, R. (2016). A Study on Crop yield Forecasting using Classification Techniques. IEEE.

Supriya, D. M. (2017). Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining approach. International Journal of Innovative Research in Computer and Communication Engineering, 5, 9648-9652.

Verma, A., Jatain, A. & Bajaj, S. (2018). Crop Yield Prediction of Wheat using Fuzzy C Means Clustering and Neural Network. International Journal of Applied Engineering Research, 13(11), 9816-9821.

Yiu. T. (2019). Understanding Random Forest. Retrieved March 30, 2020, from https://towardsdatascience.com.

***