# Agricultural Product Price And Crop Cultivation Prediction Based On Data Science Technique

**Jagadesh Waran P[1], SaranrajS[2], Mrs.R.Aiswarya[3]**
[1, 2, 3] Dept of Computer Science and Engineering
[1, 2, 3] GKMCET Chennai, Tamilnadu, India

***Abstract-*** *Crop cultivation prediction is an integral part of agriculture and is primarily based on factors such as soil, environmental features like rainfall and temperature, and the quantum of fertilizer used, particularly nitrogen and phosphorus. These factors, however, vary from region to region: consequently, farmers are unable to cultivate similar crops in every region. This is where machine learning (ML) techniques step in to help find the most suitable crops for a particular region, thus assisting farmers a great deal in crop prediction. The feature selection (FS) facet of ML is a major component in the selection of key features for a particular region and keeps the crop prediction process constantly upgraded. This work proposes a novel FS approach called modified recursive feature elimination (MRFE) to select appropriate features from a data set for crop prediction. The proposed MRFE technique selects and ranks salient features using a ranking method. The experimental results show that the MRFE method selects the most accurate features, while the bagging technique helps accurately predict a suitable crop. The performance of proposed MRFE technique is evaluated by various metrics such as accuracy (ACC), precision, recall, specificity, F1 score, area under the curve, mean absolute error, and log loss. From the performance analysis, it is justified that the MRFE technique performs well with 95% ACC than other FS methods.*

***Keywords***- Agriculture, classification, crop prediction, feature selection (FS), modified recursive feature elimination (MRFE).

## I. INTRODUCTION

AGRICULTURAL research has strengthened the economy worldwide, and is an area that offers humanity, as whole, inestimable benefits. Crop prediction in agriculture continues to present difficulties, notwithstanding current developments that include the use of an array of technological resources, tools, and procedures. Agri technology and precision farming, now termed virtual farming, have emerged as new scientific areas of interest that use data-intensive methods to boost agricultural productivity and reduce the impact on the environment. Accurately identifying crops for cultivation, based on soil and environmental factors, is critical to agricultural productivity and has been an active research topic for decades. Most of the existing approaches use machine learning (ML) for crop yield estimation, though very little has been done to predict region-specific crops based on soil and environmental parameters. Parameters such as soil type, nutrients (nitrogen, phosphorus, and potassium), micronutrients (iron, boron, and manganese), temperature, and rainfall influence crop cultiva- tion. Since the parameters differ for every zone, thus making for a massive crop prediction data set, there is a need to select crucial features that help identify suitable crops for specific areas of land.

ML algorithms play a major role in prediction. For enhanced ML performance, FS techniques [1]–[6] are used to reduce overfitting and ascertain salient features from the data set for the prediction process. The FS technique is divided into three categories: filter [7], wrapper [8], and embedded [9]. Filter methods are independent of the performance of the classifier, whereas wrapper methods select features based on its performance. The embedded method, which combines the filter and wrapper methods, is somewhat similar to the latter. This work pays special attention to wrapper FS techniques. The features selected are fed to the k-nearest neighbor (kNN), Naive Bayes (NB), decision tree (DT), support vector machine (SVM), random forest (RF), and bagging classifiers to predict a suitable crop, and evaluate the performance of the FS process. The objective of this work is to select key features from a data set and improves crop prediction performance. The main contribution of this work is to propose a novel modified recursive feature elimination (MRFE) technique to select the most appropriate key features using permutation crop data set based on soil and environmental factors, while using permutation data set, the algorithm need not to be updated with the data set at each iteration, so it reduces the computational time than existing RFE method.

### *A. Related Work*

Several studies on FS that have been undertaken for improved prediction are discussed in this section. Gregorutti et al. [10] compared the RFE and non-RFE (NRFE) techniques. The permutation importance (PIMP) mea- sure was used as a ranking criterion for FS, and the technique was tested on the

Landsat satellite data collected from the Uni- versity of California Irvine (UCI) ML repository. From the results, it was concluded that the RFE is more efficient than the NRFE. Hall and Holmes [11] compared several FS techniques and used benchmark data sets for evaluation. The results show that the wrapper technique is best for FS. Liu and Yu [12] analyzed the existing FS techniques for classification and clustering. Certain real-world applications were used in their work to demonstrate the FS techniques

Granitto et al. [13] com- pared RF-RFE with SVM-RFE. A performance evaluation was carried out using the proton transfer reaction-mass spectrome- try (PTR-MS) data of agro-industrial products. Their analysis concluded that the RF-RFE works better than the SVM-RFE. Araúzo-Azofra and Benítez [14] used 36 data sets from the UCI, Orange (Org), and silicon graphics (SGI) to evaluate mis- cellaneous FS techniques. An experimental analysis concluded that the wrapper approach is the best for selecting features.

Altmann et al. [15] proposed an improved RF model with the PIMP measure for FS. The PIMP ranking measure and Gini importance were compared to find that the PIMP-RF model significantly outperformed the Gini-RF model. Kursa and Rudnicki [16] described the Boruta FS technique, and the Boruta package provided their algorithm a convenient interface, with the Madalon data set being used for their experimental analysis. Ruß and Kruse [17] proposed a novel FS technique for wheat yield prediction with two regression models, support vector regression (SVR) and the regression tree (Reg tree), for a comparison. Darst et al. [18] compared the RF and RF-RFE in terms of the selection of variables, and concluded that the latter was not likely to scale to high-dimensional data. Hsieh et al. [19] used the RFE algorithm to select key features that impact rice blast disease (RBD). Their work analyzed climatic data collected over five years. Table I illustrates the characteristics comparison of the proposed MRFE technique with existing FS techniques such as sequential forward feature selection (SFFS), Boruta, and RFE.

### *B. Motivation and Justification*

Farming plays a critical role in the global economy, in which crop prediction is a decisive factor. FS and classification [20] are central to the crop prediction process. The literature review makes it plain that the wrapper FS technique [21]–[23] predicts crops better than existing techniques. The RFE technique is a wrapper-type FS method that works by searching for a subset of features, commencing with all features in the training data set, and thereafter successfully removing features until only a desired number remains. The RFE method ranks appropriate features in terms of their importance, discarding the least important ones. The feature that is selected impacts classification accuracy (ACC) as well. This method, however, needs an iterative process for data set updation in the feature elimination process. Updating the data set is the most difficult part of the RFE, and maximum time is taken to eliminate weak features. Motivated by these facts, this work proposes a new

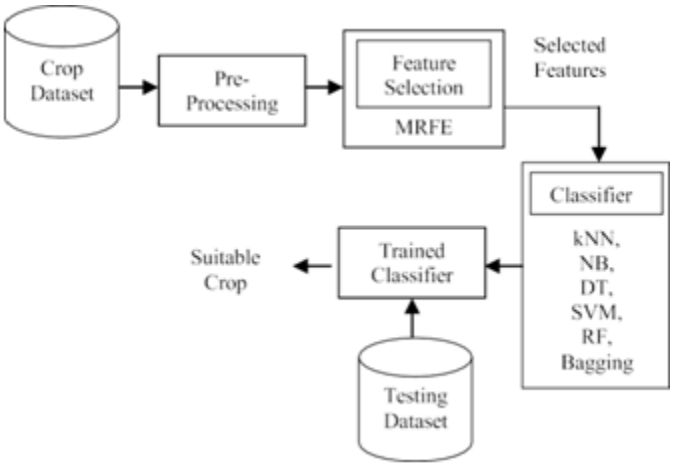


Fig.1.Outlineofthework.

FS technique called the MRFE to overcome the limitations of the RFE. The efficiency of the MRFE is analyzed, following the results of the experiments. After the features are selected, classification algorithms take the lead in the prediction process. In much of the research, a single prediction model (such as the kNN [14], NB [24], DT [25], and SVM [26]), along with an ensemble prediction model (like the RF [27]), and Bagging [28] techniques have been used to classify crop prediction. Each algorithm displays prediction characteristics of its own. However, there is a need to find the classifier that works best with the proposed FS technique for crop prediction. Therefore, this work analyzes the performance of each classifier with the proposed MRFE technique to predict the most suitable crops for specific land areas.

### *C. Outline of the Work*

Fig. 1 depicts the overall process of the proposed work. The data set containing soil and environmental features is preprocessed to find missing values and remove redundant data. The preprocessed data are then fed into the proposed MRFE FS algorithm. The features selected are input into the classifier for the learning process. This work uses a supervised learning technique for the prediction process. Training samples are trained with the classifier and unknown samples provided to validate the trained classifier. Finally, the results are evaluated, using certain performance metrics, to produce the most suitable crop.

### *D. Organization of This Article*

The remaining part of this article is organized as follows. Section II describes the existing FS techniques and the pro- posed MRFE technique. Section III discusses the existing class sification techniques to predict the suitable crop. Section IV depicts the crop prediction procedure for cultivation. Section V analyses the experimental results and Section VI concludes the work.

## II. FS TECHNIQUES

FS, which is a preprocessing step in ML [11], removes irrelevant features so as to render the classification models most efficient [24]. Sections II-A and II-B describe existing wrapper FS techniques such as SFFS, Boruta, RFE, and the proposed technique, MRFE.

### *A. Existing FS Techniques*

*1)* Sequential Forward Feature Selection: Sequential fea- ture selection (SFS) is a wrapper-based FS technique. This algorithm is divided into two, SFFS and sequential backward feature selection (SBFS). This work takes the SFFS for the FS process, the working of which is given in [29]. It starts with an empty set, selects important features from the data set, and repeats the process until every important feature is selected. The SFFS algorithm is based on the Akaike information criterion (AIC) value for FS [30].

*2)* Boruta: Boruta: The Boruta algorithm is a wrapper FS technique built around the RF classification algorithm. The advantage of RF classification is that it runs quickly and, in addition, estimates the importance of features [16]. The results provide a Z score. In the Boruta, the Z score has a great impact on the FS technique. The pseudo code of the Boruta algorithm is mentioned in [31].

*3)* Recursive Feature Elimination: The RFE is the most frequently used wrapper FS technique. The RFE starts with a whole data set and removes its weak features using a ranking method. It then updates the data set and continues the process until all the weak features are eliminated. In the RFE, the Gini importance ranking method is used for feature elimination. The pseudo code for the RFE technique is given in Algorithm 1.

### *B.* Proposed FS Technique

*1)* Modified Recursive Feature Elimination: The proposed MRFE technique removes weak features from the data set using the permutation data set and ranking method. The permutation data set shuffles the values in each field and duplicates the crop data set fed as input. Fig. 2 shows the process of the MRFE technique.

*Step 1: Initiating the Permutation Process:*

1. The given input crop data set is considered an n m matrix, where n represents records of each area and m represents features.

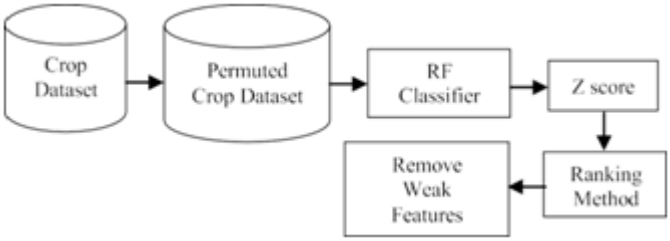


Fig.2.Flow diagram of MRFE process.

Example:Givenmatrix

| | | |
|---|---|---|
| 1 | 4 | 7 |
| 2 | 5 | 8 |
| 3 | 6 | 9 |

Shuffledmatrix

| | | |
|---|---|---|
| 4 | 1 | 7 |
| 8 | 5 | 2. |
| 3 | 6 | 9 |

This process does not affect the feature values. A data set that contains n m records shows no change following permutation application.

*2)* The shuffled data set is then combined with the input data set, i.e., the crop data set.

In the example below, the given matrix is combined with the shuffled matrix.

Combined matrix

| | | | | | |
|---|---|---|---|---|---|
| 1 | 4 | 7 | 4 | 1 | 7 |
| 2 | 5 | 8 | 8 | 5 | 2 |
| 3 | 6 | 9 | 3 | 6 | 9. |

Extending the data set results in a drop in the standard deviation value, indicating that the value is close to the mean. The permutation process offers two distinct advantages. The first is its ability to standardize the co-efficient of variables to help the ranking process eliminate weak features from the data set. The second is that the model needs no retraining as forward or backward, thus making it faster than the existing RFE technique.

*Step 2: Finding the Most Important Features:*

The RF classifier is used to discover the most important features as well as the mean decrease value that

helps find the Z score. The extended crop data set is fed into the RF classifier to find the most important features. The two main parameters of the RF classifier are as follows.

1) mtry: This refers to the number of variables that are used as each split, and is called the mtry parameter. The recommended value for the mtry is the root square of the number of features.
2) ntree: This refers to the number of trees, called the ntree parameter, which decides the splitting range of trees in the forest, with the default ntree used in the RF classifier being 100.

Step 3: Finding Z Score:

The Z score is the standard score that is used to compare the importance of the features selected. To fine-tune the performance of the RF classifier and evaluate it in this work, the ntree value is altered from 100 to 500. The basic Z score formula is given as follows:

$$Z \text{ score} = \text{mean decrease accuracy loss}(x - \mu) \div \sigma$$

where x represents the observed value, μ the mean value of the samples, and σ the standard deviation of the samples.

Step 4: Applying the Ranking Method:

Finally, a ranking method is applied to find weak soil and environmental features from the data set. Several ranking methods [32], [33] are used for FS. This work evaluates the performance of rank aggregation [34], Gini importance [27], PIMP [15], and actual impurity reduction importance (AIR-IMP) [35] to find the best ranking method for FS so as to refine the crop prediction process. The AIRIMP ranking method outperforms others and is discussed below in the section on results. Hence, it is used in the proposed MRFE FS technique to rank every feature, from the best to the worst.
2) Algorithm for MRFE: The pseudo code for the proposed MRFE technique is given in Algorithm 2.

## III. CLASSIFICATION TECHNIQUE

Classification is the learning process used in ML to predict the target class of a given input. Classification technique is divided into two, supervised and unsupervised. In this work, supervised learning methods such as the kNN, NB, DT, SVM, RF, and Bagging are used for the crop prediction process. In addition, they help evaluate the performance of the FS technique.

### A. k Nearest Neighbor

The kNN is a supervised learning process that predicts a suitable crop, based on the closest training samples, and is centered on a distance measurement for the prediction process [14]. Using the distance measurement, a new sample from the testing set is allocated to a particular target class, based on how closely it matches the training set.

### B. Naive Bayes

The NB classifier [24] is a simple classification algorithm that estimates the probability of every class and chooses a suitable crop with the maximum probability. The NB classifier is trained with the training samples, and its performance is evaluated by using testing samples from the testing set to find the most appropriate crop for cultivation. Fundamentally built on the Bayesian theorem, its principles are drawn from graph and probability theories.

### *C.* Decision Tree

The DT is a supervised learning model with a tree-like structure. Each internal node is labeled with an input feature [25] and follows a top-down approach. Each leaf node is labeled with the class used to predict the target variable [25]. For the DT, which holds the prediction class, tree splitting is important. Using the splitting, data values from the testing set are used to identify a suitable crop.

### D. Support Vector Machine

The SVM classifier is a supervised learning process that predicts the most suitable crop from the testing set. It separates classes into categories, with several possibilities for hyper plane, using the maximum margin [36]. Hyper plane, also known as decision boundary, helps classify crops. The crop that lies closest to the decision boundary is recommended for cultivation. In the SVM, finding the decision boundary is an optimization problem .

## IV. CONCLUSION

Predicting a suitable crop for cultivation is critical to agri- culture. In this work, the MRFE, a novel approach, has been proposed for selecting salient features using a permutation crop data set and a ranking method to identify the most suitable crop for a particular region. Experiments were conducted to evaluate the efficiency of the proposed MRFE technique using the kNN, NB, DT, SVM, RF, and bagging classification techniques to predict the most suitable crops for cultivation. Soil and environmental factors were considered for an analysis of the crop prediction process. The results indicate that the MRFE with the bagging technique classifier

gives better crop prediction ACC than the MRFE with other classifiers. The performance of the MRFE technique for the crop data set was assessed and compared with existing techniques like the SFFS, Boruta, and RFE. Furthermore, the suitability of the proposed MRFE technique was evaluated using three benchmark data sets. The results show that the proposed MRFE technique outperforms the others. Nevertheless, the MRFE technique needs performance-wise improvements before it can be used in large feature data sets.

## REFERENCES


[1] A. Mark Hall, "Feature selection for discrete and numeric class machine learning," Comput. Sci., Univ. Waikato, pp. 359–366, Dec. 1999.

[2] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Jan. 2003.

[3] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection–theory and algorithms," in Proc. 21st Int. Conf. Mach. Learn. (ICML), 2004, p. 43.

[4] P. S. Maya Gopal and R. Bhargavi, "Feature selection for yield prediction in boruta algorithm," Int. J. Pure Appl. Math., vol. 118, no. 22, pp. 139–144, 2018.

[5] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applica- tions," IEEE Trans. Comput. Social Syst., vol. 8, no. 1, pp. 214–226, Feb. 2021.

[6] K. Ranjini, A. Suruliandi, and S. P. Raja, "An ensemble of heterogeneous incremental classifiers for assisted reproductive technology outcome prediction," IEEE Trans. Comput. Social Syst.early access, Nov. 3, 2020, doi: 10.1109/TCSS.2020.3032640.

[7] H. Liu and R. Setiono, "A probabilistic approach to feature selection-a filter solution," in Proc. 13th Int. Conf. Int. Conf. Mach. Learn., vol. 96, 1996, pp. 319–327.

[8] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.

[9] H. Wang, M. Taghi Khoshgoftaar, and K. Gao, "Ensemble feature selection technique for software quality classification," in Proc. 22nd Int. Conf. Softw. Eng. Knowl. Eng., 2010, pp. 215–220.

[10] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and vari- able importance in random forests," Statist. Comput., vol. 27, no. 3, pp. 659–678, May 2017.

[11] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," IEEE Trans. Knowl. Data Eng., vol. 15, no. 6, pp. 1437–1447, Nov. 2003.

[12] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502, Apr. 2005.

[13] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recur- sive feature elimination with random forest for PTR-MS analysis of agroindustrial products," Chemometric Intell. Lab. Syst., vol. 83, no. 2, pp. 83–90, Sep. 2006.

[14] A. Araåzo-Azofra and J. M. Benítez, "Empirical study of feature selection methods in classification," in Proc. 8th Int. Conf. Hybrid Intell. Syst., Barcelona, Spain, Sep. 2008, pp. 584–589.

[15] A. Altmann, L. Toloåi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," Bioinformatics, vol. 26, no. 10, pp. 1340–1347, May 2010.

[16] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," J. Stat. Softw., vol. 36, no. 11, pp. 1–13, 2010.

[17] G. Ruß and R. Kruse, "Feature selection for wheat yield prediction," in Research and Development in Intelligent Systems. London, U.K.: Springer, 2010, pp. 465–478.

[18] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," BMC Genet., vol. 19, no. S1, p. 65, Sep. 2018.

[19] J.-Y. Hsieh, W. Huang, H.-T. Yang, C.-C. Lin, Y.-C. Fan, and H. Chen, "Building the rice blast Disease Prediction Model based on Machine Learning and Neural Networks," Easy Chair World Sci., vol. 1197, pp. 1–8, Dec. 2019.

[20] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," Genome Informat., vol. 13, no. 13, pp. 51–60, 2002.

[21] J. Camargo and A. Young, "Feature selection and non-linear classifiers: Effects on simultaneous motion recognition in upper limb," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 4, pp. 743–750, Apr. 2019.

[22] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta—A system for feature selection," Fundam. Inf., vol. 101, no. 4, pp. 271–285, 2010.

[23] R. Rajasheker Pullanagari, G. Kereszturi, and I. Yule, "Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regres- sion," Remote Sens., vol. 10, no. 7, pp. 1117–1130, 2018.

[24] A. Choudhary, S. Kolhe, and H. Rajkamal, "Performance Evaluation of feature selection methods for Mobile devices," Int. J. Eng. Res. Appl., vol. 3, no. 6, pp. 587–594, 2013.

[25] F. Balducci, D. Impedovo, and G. Pirlo, "Machine learning applications on agricultural datasets for smart farm enhancement," Machine, vol. 6, no. 3, pp. 38–59, 2018.

[26] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in Proc. Adv. Neural Inf. Process. Syst., vol. 13, 2001, pp. 668–674.

[27] A. Bahl et al., "Recursive feature elimination in random forest classifica- tion supports nanomaterial grouping," NanoImpact, vol. 15, Mar. 2019, Art. no. 100179.

[28] D. H. Zala and M. B. Chaudhri, "Review on use of BAGGING technique in agriculture crop yield prediction," Int. J. Sci. Res. Develop., vol. 6, no. 8, pp. 675–677, 2018.

[29] F. Shirbani and H. Soltanian Zadeh, "Fast SFFS-based algorithm for feature selection in biomedical datasets," Amirkabir Int. J. Sci. Res., vol. 45, pp. 43–56, Dec. 2013.