https://doi.org/10.48047/AFJBS.6.14.2024.8913-8918

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

CROP RECOMMENDATION SYSTEM USING HYBRID CLASSIFICATION ALGORITHM

G. Buvaanyaa[1], Dr.M.Gobi[2], Dr.R.Sridevi[3]

[1] PhD, Research Scholar, PG & Research Department of Computer Science, Chikkanna Government Arts College, Tiruppur, Tamil Nadu, India.
buvaanyaa220497@yahoo.com

[2] Associate Professor, PG & Research Department of Computer Science, Chikkanna Government Arts College, Tiruppur, Tamil Nadu, India.
gobimohan2786@gmail.com

[3] Associate Professor, Department of Computer Science, CHRIST (Deemed to be University), Bangalore, Karnataka – 560029
sridevi.r@christuniversity.in

Volume 6, Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 15 Aug 2024

doi: 10.48047/AFJBS.6.14.2024.8913-8918

Abstract: Agriculture is absolutely essential to human existence. India is the place where there is agribusiness and it is the significant wellspring of economy. Most of Indian populace straight forwardly depends on farming. So, the prediction of crops is especially important in agriculture and mostly dependent on the soil and environmental factors, such as temperature, humidity and rainfall. Prior research has achieved accurate classification through appropriate feature selection; however, the prediction of feature selection is time consuming. The Hybrid Classification Algorithm is presented by combining the Improved Recursive Feature Selection Algorithm with classification is produced. Now, this present Feature Selection methods greatly enhance the circumstances that redundant features will appear in the final subset, so that finding and removing them can improve the classification accuracy in most cases. To address this issue hybrid classification algorithm has used. Comparing this hybrid classification algorithm to other classification algorithms like Naive Bayes, Random Forest algorithms reveals that it has a high performance metrics called Precision, Recall, Accuracy and F1-Score ratio (The Predictive ability of the model).

Keywords: Data mining, Agriculture, Hybrid Classification, Feature Selection, Crop Prediction.

1. INTRODUCTION

Data mining is a technique for gathering and transforming secret data from a foundation into an understandable framework for subsequent usage. This is the computer algorithm used to search through huge datasets that combine techniques from artificial intelligence, machine learning, statistics, and database systems to find patterns [4]. Data mining, particularly statistical data mining and prediction data mining, is the final goal. To gather data from enormous datasets, numerous methods have been created over the years. This issue can be resolved in a number of ways, including grouping, the law of association, clustering, and other approaches. Utilizing the data provided by numerous identified samples, unknown samples are classified. Because it is typically used to train the use of the classification technique, this set is also known as a training set. It is possible to think of the classification task as a supervised process in which each instance is suggested to belong to a class by the significance of a particular target attribute or just by the characteristics of a class.

As datasets continue to expand in size, selecting the optimal subset of features from a crop yield raw dataset is critical to achieving the best performance in a given machine learning assignment. For creating a classification model, it is extremely critical to have an effective and constrained feature (attribute) subset. High-dimensionality datasets are prone to over fitting issues, which prevent the development of a trustworthy model. Requiring a large amount of processing power and storage space, these datasets also often lead to

models with low classification accuracy [7-11]. In feature selection, a smaller subset of pertinent characteristics is chosen from a larger original set based on predetermined criteria, such as classification performance or class separability. For machine learning applications, it is crucial. When dealing with classification problems involving tiny samples, when the number of training samples available is significantly less than the number of features, feature selection becomes very crucial.

Utilizing data cleaning methods, the data set is preprocessed. Following the removal of irrelevant features using a feature selection method, classification using machine learning techniques is applied to the data set. This aids in the type and duration of crop prediction for a certain field.

The objective of the proposed system to develop a capable improved feature selection and classification technique that works well to assess a density and manage the scale, which affects the crop prediction results. This paper presents the methodologies namely data preprocessing, Improved Recursive Feature Selection (IRFS) algorithm and Hybrid Classification Algorithm is used to discover the crop prediction results and time efficiency.

The primary contributions of the suggested method are as follows:

- The goal of developing an efficient classification method is to eliminate irrelevant features and crop examination using Hybrid classification Algorithm.
- The proposed Hybrid classification Algorithm is classification case in order to assess the prediction class. With the help of its features, the random selection feature classification achieves 22 classes.

2. LITERATURE REVIEW

According to Paja et al. (2018) [12] feature selection techniques are useful for lowering dimensionality, eliminating unnecessary data, raising learning accuracy, and enhancing result comprehensibility. But in terms of efficacy and efficiency, many of the feature selection techniques currently in use are severely challenged by the recent surge in data dimensionality. This chapter describes the major common classes type of algorithms as well as an overview of the reasons to use ranking attribute selection techniques. Also, some background information on ranking method problems is provided. Following that, the writers concentrated on a few algorithms built using rough sets and random forests. Generational Feature Elimination (GFE), a recently developed technique based on decision tree models, is also presented.

One of the most frequent issues that Indian farmers deal with, according to Reddy et al. (2019) [13], is that they choose crops without considering the soil's requirements, which causes them to experience significant production losses. Precision farming is a viable solution to this issue. Along with the benefits of improved input and output efficiency and improved farming decision-making, precision agriculture also contributes to a decrease in unsuitable crop production, which raises yield. Their approach provides solutions such as recommending a crop with high specific accuracy and efficiency based on soil parameters.

A crucial component for the survival and expansion of the Indian economy was demonstrated by Shafiulla Shariff et al. in 2022 [16]. India is a major producer of many different kinds of agricultural goods. A key component of crop cultivation is the soil. One dynamic, non-renewable natural resource that is essential to life is soil. In the past, farmers with practical experience were responsible for cultivating crops. It is no longer possible for farmers to select the ideal crop based on the qualities and features of the soil. In order to suggest the crop that can be harvested in that specific soil, a recommendation system utilizing machine learning algorithms has been built.

3. RESEARCH METHODOLOGY

Every experiment with a large-scale real-world dataset that was tested with the proposed methodology was successful. The research approach takes into account an effective classification procedure that is derived in this section and uses the Improved Recursive Feature Selection (IRFS) algorithm, data pretreatment, and Hybrid Classification Algorithm process. Figure 1 displays the overall proposed architecture diagram.

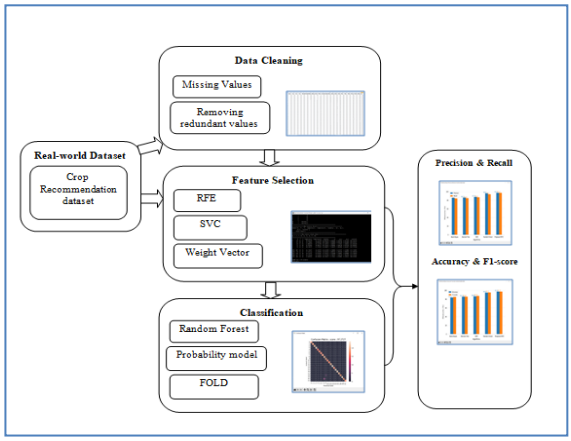

Fig. 1: Proposed Workflow Architecture Process flow

3.1 DATA COLLECTION AND PREPROCESSING

Data collection is the most effective way to gather and quantify information from many sources, such as the Kaggle website. To obtain a system's estimated dataset. This dataset has to have the following characteristics. The following factors will be taken into account for crop prediction: i) Soil PH ii) Temperature iii) Humidity iv) Rainfall v) Crop data vi) NPK.

Data cleaning is the process of data preprocessing. Preprocessing the dataset is necessary before training the model. There are several steps to data preprocessing; the first is reading the gathered dataset, which is followed by data cleaning. Some redundant attributes were removed from the datasets during data cleaning; these attributes are not taken into account when predicting crops. Therefore, in order to improve accuracy, IRFS must remove undesirable attributes and datasets that contain missing values. These missing values must be removed or filled up with unwanted NaN (Not a Number) values.

3.2 Improved Recursive Feature Selection (IRFS) Algorithm

The improved Recursive Feature Selection (IRFS) approach is evolved from the Recursive Feature Elimination (RFE) technique. By retraining the Support Vector Coefficient (SVC) on the remaining features, the proposed IRFS recursively eliminates features at each step and re-ranks the remaining features. IRFS will just eliminate a feature if it is weak. When combined with additional qualities, a poor feature could nevertheless be valuable. Thus, classification performance may suffer even from the simple removal of weak or duplicate features. When a feature is initially eliminated, IRFS evaluates the classification performance in terms of the related weight vector value in order to determine the significance of the potentially weak feature.

Let a data set Dt = $\{dt_n | n=1,2, \ldots, p\}$ incorporate p amount of data's $dn_n \in R^{nf}$; nf is the number of features $feat_i$. To denote dt^n_q as the significance of the q^{th} feature $feat_t$ of dataset d_n then every data dt_n of set Dt can be symbolized by a vector

$$D_n = \{d_n^1, d_n^2, \ldots, d_n^{nf}\} \quad \text{---------- (1)}$$

Assume that a set Feat represents all features:

$$Feat = \{feat_1, feat_2, \ldots, feat_{nf}\} \quad \text{---------- (2)}$$

When the required number of features is obtained, the weakest features (or attributes) are removed as part of the feature selection process of algorithm 1. By recursively deleting a small number of features per loop, an IRFS aims to remove any potential dependencies and co-linearity in the model. According to the model's co-efficient or feature importance dimensions, features are ranked in order of relevance. The following equation is used to compute m_c, or the total number of hyper-planes, in the event that the dataset contains more than two classes.

$$m_c = c(c - 1)/2 \quad \text{---------- (3)}$$

Equation (3) refers to the decision function and Equation (4) is for a multi-class dataset

Algorithm 1: Improved Recursive Feature Selection

Input: Dataset D, Feature Feat, Class C, Ranked Feature 0

Output: Selected feature Sf

Process

Step 1: Create logistic regression model decision using equation (3)

Step 2: For k=1 to Length(Feat) // For whole features

Perform support vector co-efficient (k) using below equation

$$A(k) = \text{sign}(k \times wt_m), m = 1,2, \ldots, m_c \quad \text{---------- (4)}$$

In the case of the linear assessment function, wt is a vector perpendicular to the hyperplane that provides a linear judgment function, and k indicates a vector containing the components of a specified spectrum.

Step 3: Initialize remaining features rfeat = rfeat – 1 and j = 1.

Step 4: When a feature's weighted vector has a high value, it indicates the feature can effectively discriminate between the classes. The weighted vector is positioned where the decision boundary has the highest margin. Utilizing Equation (5), the weight value for the assessment of variable relevance based on the Improved RFE score is determined.

$$\text{Wscore}_{\text{feat1}} = \frac{1}{m}\sum_{l=1}^{m} w_l(\text{rfeat}) + C_l \quad \text{---------- (5)}$$

Step 5: Construct a new feature set F1 by removing the feature fr from F.

Step 6: Train an Improved RFE on the samples with the remaining features and calculate the score,

$$\text{Wscore}_{\text{feat2}} = \frac{1}{m}\sum_{l=1}^{m} w_l(\text{rfeat}_1) + C_l \quad \text{---------- (6)}$$

Step 7: if Wscorefeat1 < Wscorefeat2

RF = RF ∪ feat_m = 1

Else if j = rfeat

Remove the primary feature

RF = RF ∪ f1, rfeat= rfeat –feat1

Else

m = m+1;

Step 8: Repeat Steps 1 to 7 until rfeat = 0

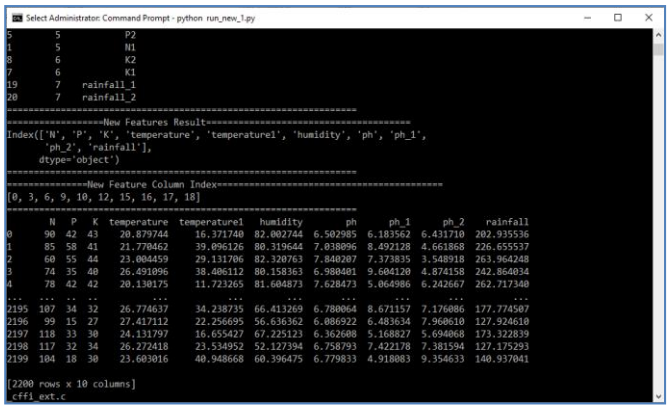

Fig. 2: Selected Features

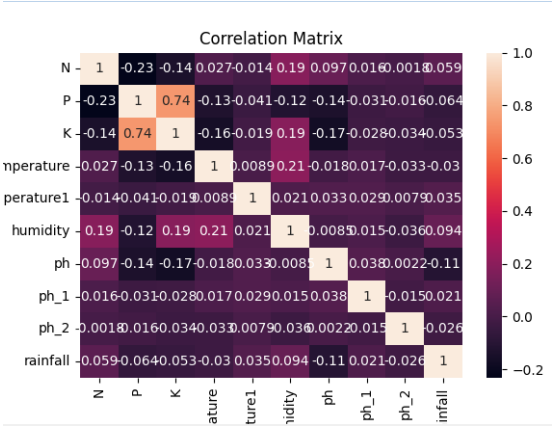

Fig. 3: Selected Correlation Feature matrix

3.3 Hybrid Classification Algorithm

In this research, a new method for Hybrid Classification Algorithm, an expansion of the Random Forest and Probability Classification Model, is introduced. The primary step in this classification is to initialize the number of tree variables and dimensions. The inputs for classification are the final feature choice as a test feature and every preprocessed feature as a training feature. With this method, the feature set for the feature dimension is obtained by first processing the training and testing features. The advantage of Hybrid Classification method is capable of effectively handling high-dimensional datasets. Compared to the random forest method, the Hybrid Classification algorithm is more accurate at forecasting outcomes.

A variety of FOLD (First Order Logical Decision) methods are used to classify a Hybrid classification case in order to assess the prediction class. With the help of its features, the random selection feature classification achieves 22 classes.

$$S_f = \sum_{k=1}^{c}\sum_{j=1}^{\text{feat}} \text{find}(U_{\text{feat}} == \text{pred}_c) \quad \text{---------- (7)}$$

Where feat is features; c is amount of class; U denotes the trained dataset's unique features. The classification accuracy is under equation 8 following the selection feature process and Hybrid Classification Algorithm procedure.

$$\text{Classificaiton}_{\text{acc}} = \sum_{m=1}^{\text{len}(U)} \text{SF}(\text{UIndex}_m) \quad \text{---------- (8)}$$

Algorithm 2: Hybrid Classification Algorithm

Input: Training set Tr := $(x_1, y_1), \ldots, (x_n, y_n)$, features F, and amount of trees in forest FB

Output: Prediction Accuracy

Begin Process

Step 1: Initialization Pd = 0 // Pd is prediction

Step 2: For k = 1: FB

TR (k) = A sample from feature

Pd(i) = Learn(TR(k), F)

PD = PD ∪ {pd(i)} using eqn. (8)

PA ← PD

End for

Function Learn(TR, F)

At every Node:

F ← a small subset of F

Calculate selection unique feature using eqn, (7)

return prediction tree

end function

4. EXPERIMENTAL RESULT

Using Python 3.8 version simulation, it is seen through experimental assessment that the suggested Hybrid Classification algorithm performs well when run on an Intel I5 series 3.21 GHz, x64-based CPU, with 8 GB memory, and Windows 10 operating system. To compare the proposed Hybrid Classification Algorithm performance with existing Naive Bayes and Random Forest [17-18] algorithms was performed. Real-world high-dimensional datasets are attainable in this experiment. The proposed classification confusion matrix diagram is shown in figure 4. The proposed Hybrid Algorithm technique was used to calculate dataset Precision, Recall, and accuracy measures. The comparison of Precision, Recall, Accuracy and F1 - Score measure described in Table 1 and figure 5.

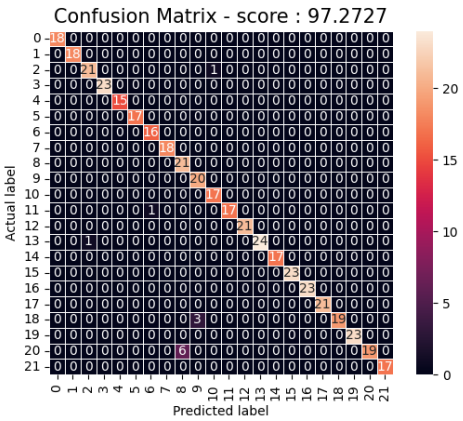

Fig. 4: Confusion Matrix

Table 1: Comparison of Crop Recommendation Datasets

Methods	**Precision**	**Recall**	**Accuracy**	**F1-socre**
Naive Bayes	85.93	84.21	83.93	85.06
Random Forest	96.47	94.00	94.91	95.21
Proposed Hybrid Classification Algorithm	**97.25**	**97.27**	**97.28**	**97.25**

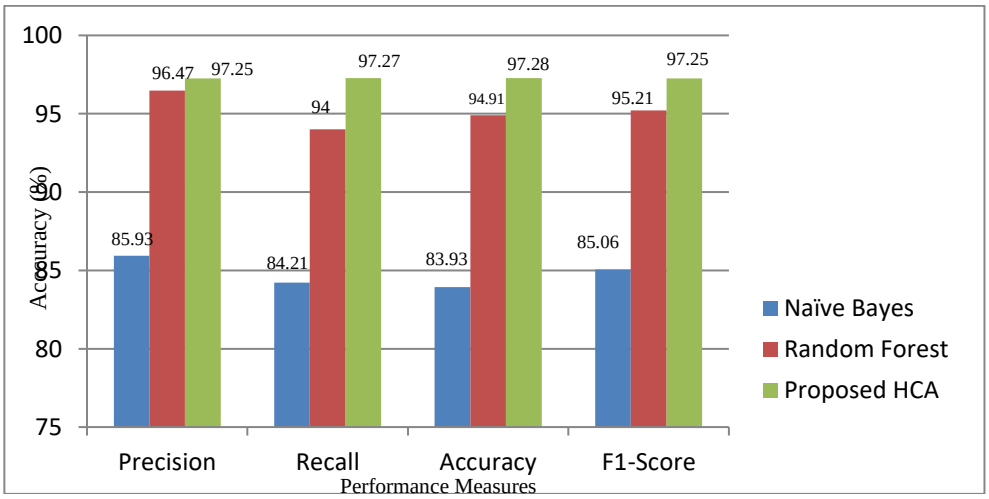

Fig. 5: Precision, Recall, Accuracy, F1-Score chart

5. CONCLUSION

In the areas of data cleaning, IRFS feature selection, and Hybrid classification Algorithm, the research examined and developed information on crop prediction discovery. Data cleaning is used to perform preprocessing on the crop recommendation datasets. The IRFS approach is used to discard the irrelevant features with feature weight score from the proposed feature selection. The proposed Hybrid Classification algorithm combines Random forest with probability classification algorithm to automatically partition feature samples based on the obtained classification. The findings show that compared to the existing classification algorithms, the suggested technique provides greater prediction accuracy.

6. REFERENCES

[1] A. Nigam, S. Garg, A. Agrawal and P. Agrawal, "Crop Yield Predict ion Using Machine Learning Algorithms," 2019 Fifth International Conference on Image Information Processing (ICIIP), 2019, pp. 125-130.

[2] N.Gandhi and L.J. Armstrong, "Applying data mining techniques to predict yield of rice in Humid Subtropical Climatic Zone of India", Proceedings of the 10th INDIACom-2016, 3rd 2016 IEEE International Conference on Computing for Sustainable Global Development, New Delhi, India, 16th to 18th March 2016.

[3] S. K. Honawad, S. S. Chinchali, K. Pawar, and P. Deshpande, "Soil classication and suitable crop prediction," in Proc. Nat. Conf. Comput. Biol., Commun., Data Anal. 2017, pp. 25-29.

[4] P. Priya, U. Muthaiah, and M. M. Balamurugan, "Predicting yield of the crop using a machine learning algorithm," Int. J. Eng. Sci. Res. Technol., vol. 7, pp. 1-7, Apr. 2018.

[5] A. Chougule, V. Kumar, and D. Mukhopadhyay, "Crop suitability and fertilizer recommendation using data mining techniques," in Progress in Advanced Computing and Intelligent Engineering (Advances in Intelligent Systems and Computing), vol. 714. Singapore: Springer, 2019.

[6] W. Paja, K. Pancerz, and P. Grochowalski, "Generational feature elimination and some other ranking feature selection methods," in Advances in Feature Selection for Data and Pattern Recognition, vol. 138. Cham, Switzerland: Springer, 2018, pp. 97-112.

[7] D. A. Reddy, B. Dadore, and A. Watekar, "Crop recommendation system to maximize crop yield in ramtek region using machine learning," Int. J. Sci. Res. Sci. Technol., vol. 6, no. 1, pp. 485-489, Feb. 2019.

[8] Shafiulla Shariff, R B Shwetha, O G Ramya, H Pushpa and K R Pooja, "Crop Recommendation using Machine Learning Techniques", Volume&Issue-ICEI-2022, vol. 10, no. 11.

[9] Prashant Kumar; Keshav Bhagat; Kusum Lata; Sushant Jhingran, "Crop recommendation using machine learning algorithms", IEEE International Conference on Disruptive Technologies (ICDT), 2023.

[10] S. P. Raja, Barbara Sawicka, Zoran Stamenkovic and G. Mariammal, "Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers", IEEE Access, 2022.

[11] Dr. S. Selvi, Dr. M. Gobi, "Improving Cloud Data Security using Hyper Elliptical Curve Cryptography & Stenography", Int. J. Scientific Research & Development, Vol. 5, Issue 4, 2017.

[12] S. Selvi, M. Gobi, "Hyper Elliptic Curve based Homomorphic Encryption Scheme for Cloud Data Security", Int. Con. On Intelligent Data Communication Technologies and Internet of Things (ICICI), 2018.

[13] S. Selvi, M. Gobi, "Hyper Elliptic Curve Cryptography in Multi Cloud – Security using DNA (genetic) techniques", Int. Con. On computing Methodologies and Communication (ICCMC), July 2017.

[14] Dr. S. Selvi, Dr. M. Gobi, "An Efficient Data Security Model using Hyper Elliptic Curve Cryptography and Stenography", Int. J. Research and Development in Technology, Vol. 7, Issue 6, June – 2017.

[15] Dr. M. Gobi, B. Arunapriya, "A Survey on public–key and Identify–Based Encryption Scheme with Equality Testing over Encrypted Data in Cloud Computing", Journal of algebraic Statistics, Vol.13,No.2,2022, p.2129-2134.

[16] Dr. M. Gobi, R. Sridevi, "ECC Encryption and LSB Data Embedding Technique for Message Security", Int. J. for Res. In Technological Studies, Vol.2, Issue 7, June 2015.

[17] G. Siva Brindha, Dr. M. Gobi, "CryptoData MR: Enhancing the Data Protection using Cryptographic Hash and Encryption/Decryption Through Map Reduce Programming Model", International Conference on Innovative Computing and Communcations (ICICC) – 2023.

[18] M. Gobi, G. Buvaanyaa, "An Efficient Naïve Bayes Imputation Method for Missing Values", IRJMETS, Vol.2, Issue 7, July-2020.

[19] R. S. Vindan, M. Gobi, "Pinning based Energy aware Computation Offloading for Mobile Cloud Computing", First Int. Con. Computational Science and Technology (ICCST), 2022.

[20] G. Siva Brindha, Dr. M. Gobi, "Improving the Map Reduce Performance using Symmetric Key Algorithm", International Journal of Science and Research (IJSR), Vol. 10, Issue 4, April 2021.