2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-3701-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/Confluence52989.2022.9734173

Smart Crop Recommender System- A Machine Learning Approach

1st Rakesh Kumar Ray
Department of Computer Science and Engineering
Centurion University of Technology & Management
Bhubaneswar, India, 752050
rakeshrayk@gmail.com
0000-0002-9374-2972

2nd Saneev Kumar Das
Department of Computer Science and Engineering
Centurion University of Technology & Management
Bhubaneswar, India, 752050
saneevdas.061995@gmail.com
0000-0002-0097-5102

3rd Sujata Chakravarty
Department of Computer Science and Engineering
Centurion University of Technology & Management
Bhubaneswar, India, 752050
chakravartys69@gmail.com
0000-0002-1293-5378

***Abstract*—Machine learning has proven its efficacy in solving agricultural problems in the recent years such as crop recommendation, crop yield prediction, and many such. With the advancement in the sub-domain of machine learning i.e., deep learning, multiple problems are minutely solved in agricultural sector. This paper focuses on recommending 22 types of crops with the aid of correlation analysis, distribution analysis, ensembling, and majority voting. A three-tiered framework is proposed in order to implement the crop recommendation problem. It includes data preprocessing, classification, and performance evaluation modules. The feature analysis is done through correlation plots and density distribution followed by classification using ensembling techniques. Finally, performance evaluation is performed using majority voting technique. This article further uses ensembling with base learners i.e., decision trees, random forest, Naïve Bayes, and support vector machines using majority voting. Further, majority voting is used to decide the final performance metrics. The practical visualization of the correlation plot, density-histogram distribution plots, confusion matrices, and performance plot are presented. The accuracy achieved post implementation is 99.54% by using Naïve Bayes classifier. The majority voting ensembler has not shown much accuracy i.e., 98.52%. Thus, Naïve Bayes classifier is proved to be the best fit for this problem statement. Some challenges and future research directions are also epitomized in this article.**

***Index Terms*—Crop Recommendation, Correlation Analysis, Kernel Density Estimation, Classification, Ensembling, Majority Voting.**

I. Introduction

With the advent of machine learning in diverse domains, an immense applicability is recently seen in the domain of precision agriculture. The economic conditions of the farmers are not usually stable and the prime reason lies in the selection of crop to grow in their respective fields. The financial losses they face are generally due to wrong selection of crop [1]. Thus, in order to tackle this problem, the interference of machine learning techniques in this domain is required highly. Recommending suitable crops is an unprecedented event and to map this using classification techniques is a hectic task [2]. Factors impacting the recommendation of crops are mainly soil characteristics, macro-nutrient composition, micro-nutrient composition, pH value, humidity, rainfall, temperature, and many such [3], [4]. Also, machine learning nowadays is competent enough to perform multi-class classification, multi-label classification, multivariate regression, and many such advanced tasks. Many classification algorithms exist like decision trees, k-nearest neighbors, logistic regression, support vector machines, AdaBoost, XGBoost and many such which can proficiently perform multi-class classification. There exists few optimum values for each parameter in order to predict the suitable crop [5]. Also, crop yield prediction is a task which is in rage nowadays in integration with remote sensing techniques.

This paper focuses on crop recommendation based on 22 types of crops which are being recommended. Thus, a multi-class classification approach is used to solve the problem. Further, the proposed model presents a generalized approach to use ensembling using majority voting technique to perform recommendation tasks. This paper initially presents few research works in this domain. The proposed framework is practically realized through diverse analyses like correlation analysis, kernel density estimation (KDE) based density distribution, classification with the aid of ensembling using majority voting, and performance evaluation.

II. Background

In this section, the enabling technologies which are used to implement the proposed framework are scrutinized in detail.

A. Correlation Analysis and Distribution Plotting

Correlation analysis is performed in order to map the correlation of each attribute with itself and all other attributes as well. Prior to performing classification task, correlation

978-1-6654-3701-1/22/$31.00 ©2022 IEEE

analysis and distribution plotting can be done to identify the features which can be removed in turn improvising the performance of the final model. The correlation plot presents how each attribute is correlated with the other. The Pearson Product Moment correlation coefficient is used to estimate the correlation between the attributes. If the correlation coefficient is positive, there exists a strong association and if the correlation coefficient is negative, there exists a weak association.

Distribution plotting is one of the major tasks which needs to be performed prior to classification. There exists multiple distribution functions i.e., probability density function, cumulative density function, parametric density function, and non-parametric density function which can also be called as kernel density function [6]. The estimation technique used in this work is non-parametric i.e., kernel density estimation technique [7]. The kernel used generally smoothens the probability of a range of outcomes. The kernel can also have different types out of which Gaussian kernel is used to perform the density estimation of each parameter.

B. Classification using Ensemblers

Classification is a supervised learning task where there exists a label which acts as the response variable and certain predictors which aid in estimating the response. Diverse classification algorithms exist but to find the best fit is a hectic task. Thus, ensembling using majority voting is a right option to perform classification tasks [8]. Ensembling is the process to combine diverse classification models to work in order to achieve a higher accuracy [9]. It also helps prevent overfitting which is a major challenge faced in machine learning. Majority voting is an ensembling technique where multiple base learners are used in order to generate the performance parameters such as accuracy, precision, recall, and F1-score [10].

III. Related Works

Dash et al. [11] presented the inevitability of micro-nutrients, macro-nutrients, and weather conditions in predicting the suitable crop that can be grown on that soil. The authors used support vector machines and decision tree classifier in order to predict the suitable crop. Also, the authors with the assistance of curve fitting and regression analysis presented the prediction of crops. Further, the paper entailed an android based framework to leverage IoT based smart agriculture system. Certain correlation patterns were plotted in order to identify features that contribute less towards classification. Further, the relation between the nutrients i.e., nitrogen (N), phosphorus (P), and potassium (K) were presented diagrammatically. The authors also visualized the user interface of the android based IoT smart agriculture application. The highest accuracy was obtained with the use of support vector machine integrated with the kernel i.e., 92%. The predicted crops included rice, wheat, and sugarcane.

Doshi et al. [12] devised an intelligent crop recommendation system with the assistance of machine learning algorithms and named it as *AgroConsultant*. The authors considered few inevitable parameters including demographic features, soil characteristics, environmental parameters, and season-based features to predict the type of crop that can be sown. The proposed architecture included two sub-systems where the former focused on recommending suitable crop based on soil and environment characteristics and integrating it with cartographic visualization. Further, the latter focused on weather-based parameters which after training of the model was integrated with monthly rainfall predictor. The authors used decision tree, k-nearest neighbor, random forest, and neural networks in order to classify the crop types. The accuracies obtained were 90.20%, 89.78%, 90.43%, and 91% respectively. The authors stated that neural network was found to provide the best fit. The authors with the use of neural networks at the backend designed an application to predict crop types whose user interface was displayed in this article. Also, the map presenting the location based crop prediction was also visualized.

Pudumalar et al. [13] proposed a novel crop recommender system which used ensembling technique with the integration of majority voting technique. The classifiers used in this work included random forest, k-nearest neighbor, CHAID, and Naive Bayes. The number of class labels included in the considered dataset was ten i.e., ten types of crop were recommended. The proposed framework included preprocessing of the dataset, feature extraction, model training using ensembling approach, and validation. The prediction accuracy obtained was 88%. The authors also shared the user interface of the developed crop recommender system.

Kulkarni et al. [14] with the aid of ensembling technique using majority voting predicted suitable crops based on parameters like surface temperature, average rainfall, crop type (i.e., either Kharif or Rabi) and certain chemical parameters of the soil. The base learners used in this article included linear support vector machines, random forest, and Naive Bayes. The considered dataset included significant parameters to predict the suitable crop. The dataset consisted of 9,000 samples out of which post train-test split, the authors obtained 6,750 samples for training set and 2,750 samples for testing set. The combined average accuracy achieved with the use of ensembling techique was 99.91%.

Garanayak et al. [15] proposed an agricultural recommendation system using regression-based approaches. The paper had a prime focus on predicting yield of diverse crops. The algorithms used in this article includes decision tree, random forest, linear regression, polynomial regression, and support vector machines. The dataset comprised of five diverse crops i.e., rice, gram, ragi, onion, and potato. Climatic factors impacting the crop yield were chosen to be vapor pressure, area, cloud cover, and season. The authors also used majority voting in order select the best fit. Plots signifying area versus production were also epitomized in a crop-based fashion for diverse selected machine learning algorithms. The accuracy achieved through majority voting in this article was 94.78%.

IV. Materials and Methods

A. Dataset Specification

The dataset is collected from [16]. The dataset comprises of 2,200 number of rows (i.e., samples) and 8 number of rows (i.e., 8 attributes). The total number of instances is 17,600. The number of predictors is 7 and the response column comprises the name of the predicted crop. The dataset comprises of the following predictors i.e., as follows:

- Nitrogen content of the soil.
- Phosphorus content of the soil.
- Potassium content of the soil.
- Soil Temperature.
- Humidity content of the soil.
- pH of the soil.
- Average rainfall in millimeters.

The crops which are being recommended are described in Table I.

TABLE I: Crops being recommended and its corresponding number of instances in the dataset.

Sl. No.	Crop	Instances
1	Pigeonpeas	100
2	Apple	100
3	Blackgram	100
4	Muskmelon	100
5	Pomegranate	100
6	Orange	100
7	Mungbean	100
8	Watermelon	100
9	Mango	100
10	Chickpea	100
11	Maize	100
12	Rice	100
13	Papaya	100
14	Grapes	100
15	Coffee	100
16	Lentil	100
17	Coconut	100
18	Cotton	100
19	Kidneybeans	100
20	Jute	100
21	Banana	100
22	Mothbeans	100

B. Proposed Framework

A generalized three-tiered architecture is proposed which is validated through proper implementation on crop recommendation application. The first module entails data preprocessing and feature extraction. Initially, the data is checked for missing values and noisy instances which needs to be handled. Further, data can be reduced by removal of unwanted instance. A correlation analysis needs to be performed in order to identify the inevitability of each attribute in the data. Further, feature extraction is possible through diverse techniques like principal component analysis, linear discriminant analysis, or t-stochastic neighbor embedding based on the type of task i.e., either classification, regression, or clustering. The second module comprises of analyzing the data and performing classification tasks. In this module, initially train-test split can be performed in order to separate training and testing instances. It can be done through k-fold cross validation, holdout validation, or percentage split. Further, ensembling techniques have shown their proficiency in diverse domains. Thus, ensembling can be done using multiple classifiers like decision tree, random forest, Naïve Bayes, and support vector machines or any other combination of classifiers. Finally, the third module i.e., model testing and validation is performed through visualization of confusion matrix of each base learner followed by majority voting technique which can select the best fit with the use of diverse performance metrics i.e., accuracy, precision, recall, and F1-score.

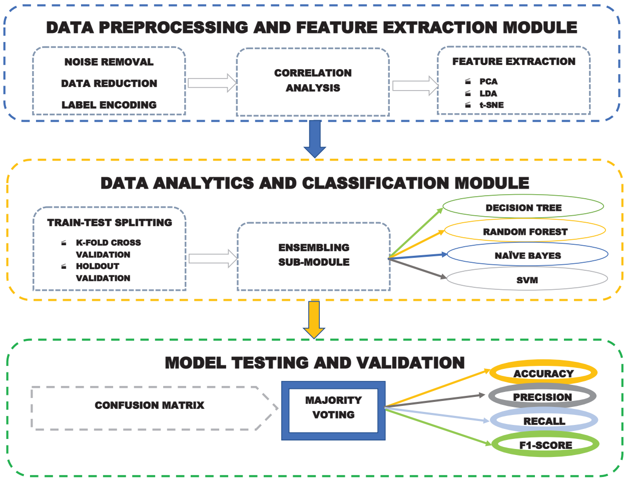
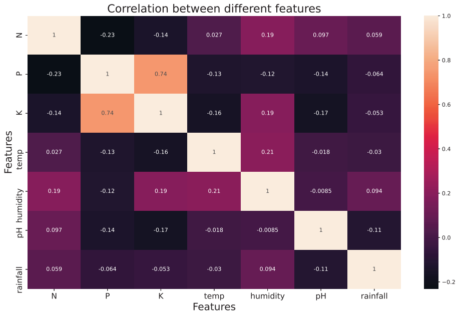


Fig. 1: Proposed framework to implement suitable crop recommendation task.

V. Results and Discussion

Our implementation leverages the multi-class classification problem. The total number of classes in the dataset is 22. Ensembling combined with majority voting technique is opted so as to perform the classification.

This section presents the practical realization of the proposed framework over the crop recommendation dataset. The considered dataset was initially imported and checked for any missing values. The data since not a real-time data, contains very less noise and thus the accuracy achieved is also high. Not much data preprocessing was required for the considered dataset. Since, the class label was in string format, label encoding technique was used so as to perform classification operations. Further, the correlation plot is visualized which is presented in Fig. 2. The correlation plot that was obtained states that there is no such attribute which can be removed since all the parameters in the dataset contribute towards the task of prediction. Further, the density plots combined with the corresponding histogram were visualized for distribution for each attribute. The distribution of nitrogen as well as phosphorus attributes are epitomized in Fig. 3. Similarly, the distribution of potassium as well as temperature can be seen in Fig. 4. Also, the distribution of the remaining three attributes i.e., rainfall, humidity, and pH level is presented in Fig. 5. In order to plot the distribution of each attribute, kernel density estimation technique is used which is a non-parametric density plotting method. The Gaussian kernel function was selected and the bandwidth was based on histogram bin widths in order to plot the distribution function for each parameter. The classification module is then implemented with the use of ensembling and majority voting approach. The confusion matrix for each base learner is presented. The selected classifiers were decision tree (Fig. 6), random forest (Fig. 7), Naï̈ve Bayes (Fig. 8), and support vector machines (Fig. 9). Finally, in order to test the model and perform majority voting, the performance metrics selected for analysis were accuracy, precision, recall, and F1-score which is plotted using a three-dimensional plot which can be seen in Fig. 10. The performance metrics for the considered classifiers is also presented in a tabular format in Table II.



	N	P	K	temp	humidity	pH	rainfall
N	1	-0.23	-0.14	0.027	0.19	0.097	0.059
P	-0.23	1	0.74	-0.13	-0.12	-0.14	-0.064
K	-0.14	0.74	1	-0.16	0.19	-0.17	-0.053
temp	0.027	-0.13	-0.16	1	0.21	-0.018	-0.03
humidity	0.19	-0.12	0.19	0.21	1	-0.0085	0.094
pH	0.097	-0.14	-0.17	-0.018	-0.0085	1	-0.11
rainfall	0.059	-0.064	-0.053	-0.03	0.094	-0.11	1

Fig. 2: Correlation plot for all the predictors in the dataset.

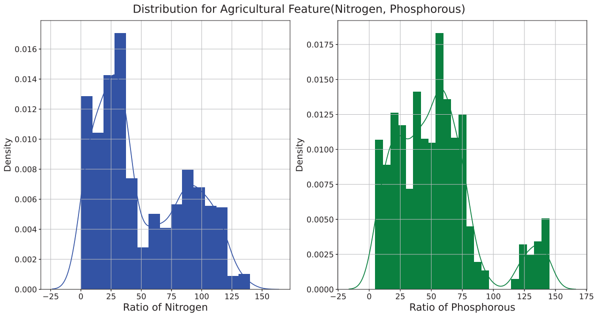


Fig. 3: KDE based density-histogram distribution plot the attributes nitrogen, phosphorus.

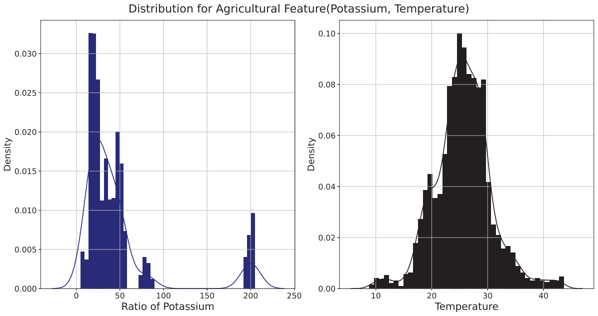


Fig. 4: KDE based density-histogram distribution plot for the attributes potassium, temperature.

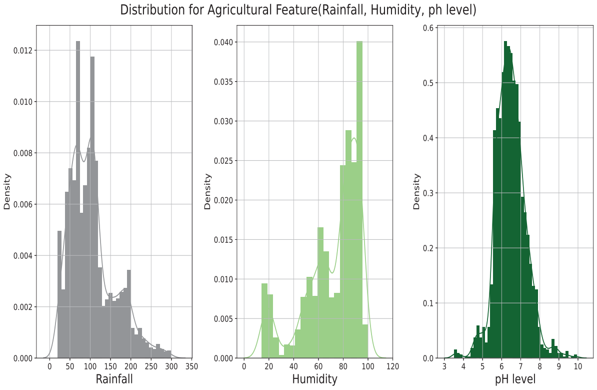
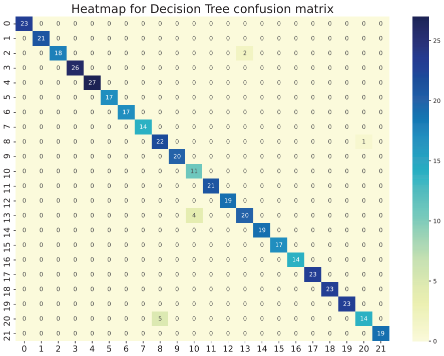


Fig. 5: KDE based density-histogram distribution plot for the attributes rainfall, humidity, and pH level.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	18	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
3	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	4	0	0	20	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0
20	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	14	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19

Fig. 6: Confusion matrix obtained for decision tree classifier.

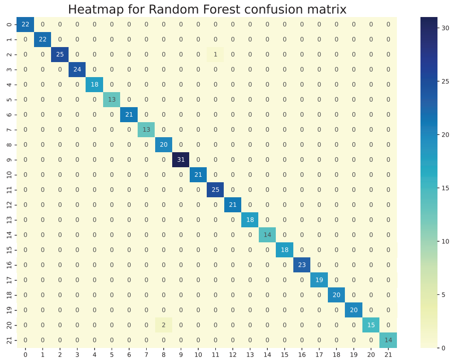


Fig. 7: Confusion matrix obtained for random forest classifier.

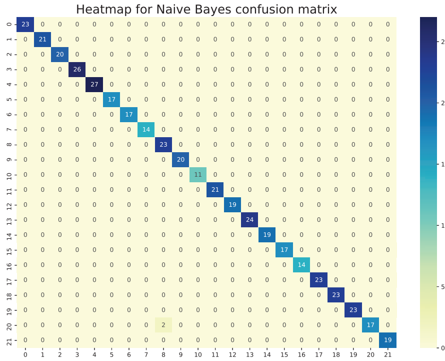


Fig. 8: Confusion matrix obtained for Naï̈ve Bayes classifier.

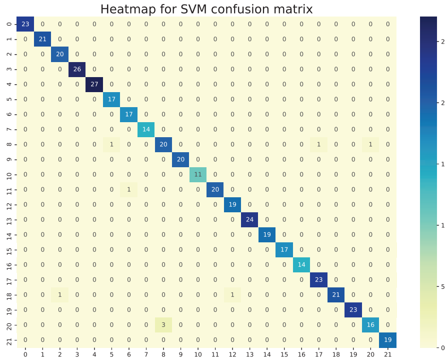


Fig. 9: Confusion matrix obtained for support vector machine classifier.

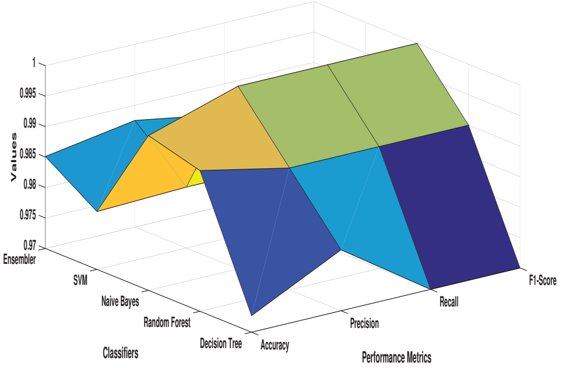


Fig. 10: A three-dimensional plot epitomizing performance parameters for all base learners and the ensembler.

TABLE II: Dataset description for three created datasets for the spatial database.

Classifiers	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Tree	97.27%	98.00%	97.00%	97.00%
Random Forest	99.31%	99.00%	99.00%	99.00%
Naï̈ve Bayes	99.54%	100.00%	100.00%	100.00%
SVM	97.95%	98.00%	98.00%	98.00%
Ensembler	98.52%	98.75%	98.50%	98.50%

VI. Concluding Remarks

The convergence of agricultural problem statements with machine learning has led towards diverse applicability. One of those is the recommendation of suitable crops based on certain inevitable parameters. This paper focused on recommending 22 types of crops with the aid of ensembling techniques using majority voting. The proposed framework was three-tiered and it entailed data preprocessing and feature extraction as a module followed by classification as a module followed by performance evaluation. The proposed model was practically realized and the visualization for each module was presented in detail. Initially, the paper epitomized correlation plot, kernel density estimation based Gaussian distribution for each predictor. Further, the confusion matrix for each base learner was presented followed by a three-dimensional plot signifying accuracy, precision, recall, and F1-score for all the base learners and the ensembler which used majority voting. Also, the tabular representation of the performance matrix was presented. Further, independently the best classifier was found to be Naï̈ve Bayes with an accuracy of 99.54%. The ensembler using majority voting produced an accuracy of 98.52%. Thus, the best fit was found to be Naï̈ve Bayes classifier. Certain future research directions exist in this domain of work where deep neural networks can be integrated so as to produce better performance [17]. Further, real time application can also be designed keeping these classifiers at the backend in order to predict the suitable crop based on the parameters provided by the user. Also, cloud based implementation using latest techniques of serverless computing in integration with location-based prediction with the aid of geospatial analysis can act as a future research direction [18]. Further, crop yield prediction, disease detection in crops, macro-nutrient recommendation can also be integrated with the crop recommendation module so as to provide higher predictive competency.

References

[1] A. Chlingaryan, S. Sukkarieh, and B. Whelan, “Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review,” *Computers and electronics in agriculture*, vol. 151, pp. 61–69, 2018.

[2] K. Patel and H. B. Patel, “A state-of-the-art survey on recommendation system and prospective extensions,” *Computers and Electronics in Agriculture*, vol. 178, p. 105779, 2020.

[3] A. Kumar, S. Sarkar, and C. Pradhan, “Recommendation system for crop identification and pest control technique in agriculture,” in *2019 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2019, pp. 0185–0189.

[4] J. Lacasta, F. J. Lopez-Pellicer, B. Espejo-García, J. Nogueras-Iso, and F. J. Zarazaga-Soria, “Agricultural recommendation system for crop protection,” *Computers and Electronics in Agriculture*, vol. 152, pp. 82–89, 2018.

[5] A. Chougule, V. K. Jha, and D. Mukhopadhyay, “Crop suitability and fertilizers recommendation using data mining techniques,” in *Progress in Advanced Computing and Intelligent Engineering*. Springer, 2019, pp. 205–213.

[6] Z. Ghahramani, “Probabilistic machine learning and artificial intelligence,” *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[7] M. C. Jones, “The performance of kernel density functions in kernel distribution function estimation,” *Statistics & Probability Letters*, vol. 9, no. 2, pp. 129–132, 1990.

[8] N. Jamali and C. Sammut, “Majority voting: Material classification by tactile sensing using surface texture,” *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 508–521, 2011.

[9] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu, and J. Li, “Gesture recognition by instantaneous surface emg images,” *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.

[10] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, “Credit card fraud detection using adaboost and majority voting,” *IEEE access*, vol. 6, pp. 14 277–14 284, 2018.

[11] R. Dash, D. K. Dash, and G. Biswal, “Classification of crop based on macronutrients and weather data using machine learning techniques,” *Results in Engineering*, vol. 9, p. 100203, 2021.

[12] Z. Doshi, S. Nadkarni, R. Agrawal, and N. Shah, “Agroconsultant: Intelligent crop recommendation system using machine learning algorithms,” in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, 2018, pp. 1–6.

[13] S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika, and J. Nisha, “Crop recommendation system for precision agriculture,” in *2016 Eighth International Conference on Advanced Computing (ICoAC)*. IEEE, 2017, pp. 32–36.

[14] N. H. Kulkarni, G. Srinivasan, B. Sagar, and N. Cauvery, “Improving crop productivity through a crop recommendation system using ensembling technique,” in *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*. IEEE, 2018, pp. 114–119.

[15] M. Garanayak, G. Sahu, S. N. Mohanty, and A. K. Jagadev, “Agricultural recommendation system for crops using different machine learning regression methods,” *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, vol. 12, no. 1, pp. 1–20, 2021.

[16] M. Ache, “Malware traffic analysis knowledge dataset 2019 (mta-kdd-19),” 2020, retrieved from, https://www.kaggle.com/atharvaingle/crop-recommendation-dataset on 25 July, 2021.

[17] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] S. Bebortta, S. K. Das, M. Kandpal, R. K. Barik, and H. Dubey, “Geospatial serverless computing: Architectures, tools and future directions,” *ISPRS International Journal of Geo-Information*, vol. 9, no. 5, p. 311, 2020.