Crop Prediction Using Feature Selection And Ensemble Techniques

S Neelufar[1] , A.P. Siva Kumar[2]

[1]M.Tech Scholar
Department of CSE (CS)
JNTUA College of Engineering
Ananthapuramu, A.P, India
neelufar1818@gmail.com
[2]Professor Department Of Cse Jntua
College of Engineering Ananthapuramu, India
sivakumar.cse@jntua.ac.in

Abstract

Research in agriculture is expanding. Agriculture relies heavily on environmental and soil aspects, including temperature, humidity, and rainfall to anticipate crops. In the past, farmers had control over the selection of the crop to be grown, monitoring the development and timing of its harvest. The difficult process of forecasting crops in agriculture has resulted in the creation and testing of several models. such as Classification Techniques of Machine learning. The purpose of this research is to enhance the accuracy of the crop forecast by employing Ensemble Techniques. Ensembling In comparison to the current classification techniques, the Decision Tree, Support Vector Machine, and Random Forest algorithms perform better and provide greater accuracy.

keywords: Crop Prediction, Ensemble Techniques, Support vector machines, Decision trees, Random forest and Feature selection.

I. INTRODUCTION

A lot of models have been developed and tested as crop forecasting in agriculture is a challenging task. As crop production depends The project calls for the use of various datasets, both on biotic and abiotic aspects. The elements of the environment known as "biotic factors" arise as the outcome of interactions between living organisms, either directly or indirectly. (Microorganisms, plants, animals,

parasites, predators, and pests). variables that are caused by humans, such as soils, irrigation, fertilisation, plant protection, and pollution of the air and water. are also included in this category. These substances may cause internal faults, structural problems, alterations chemical make-up of the crop yield, and other variances in crop output. The ecology, as well as the quality and quantity of the crop, are all shaped by the changes in the area. Abiotic Physical, chemical, and other components can all be categorised into these three groups. The recognised physical factors are soil type, geography, soil granularity, climate, and water chemistry, notably salinity. Additionally covered are climatic variables, radiation (such as ionising, electromagnetic, ultraviolet, and infrared), and mechanical vibrations (vibration, noise). Priority environmental pollutants include substances such as lead, PAHs, nitrogen fertilisers, pesticides, fluorine, sulphur dioxide, cadmium, and their derivatives, as well as nitrogen oxides and their derivatives, as well as carbon monoxide. These pollutants can all be dangerous. The others are asbestos, aflatoxins, dioxins and furans, mercury, arsenic, and so on. In addition to circumstances related to bedrock, relief, weather, and water, abiotic elements also have an impact on a substance's characteristics. The creation of soils and their importance for agriculture is influenced by a variety of soil-forming variables. The purpose the purpose of this study is to examine the potential applications of machine learning techniques. utilised in agriculture to predict crops. Due to the frequent changes in the environment, farmers now find it difficult to decide which crop to grow, to follow its development, and to anticipate when it will be ready for harvest. Therefore, prediction has been replaced by machine learning approaches. Using a range of feature selection techniques and ensemble methodologies, the study's primary objective is to pre-process the raw data into a dataset that is Machine Learning friendly. Through the use of ensemble approaches, crop prediction is accomplished. It improves accuracy, precision, memory, and F1 scores. In terms of agricultural yield prediction, ensemble techniques are the most accurate.

II. RELATED WORK

Singaraju Jyothi et al. [1] proposed an abundance of data thanks to technological advancements in computers and information storage. Since it has proven challenging to learn

anything from this raw data, many approaches and techniques like data mining that can close the knowledge gap. This study aimed to investigate whether fresh data mining approaches could identify meaningful connections in a collection of soil science data. In Tirupati, the Department of Soil Sciences and the Department of Agricultural Chemistry at S V Agricultural College have gathered a significant amount of data on soil profile measurements from several sites close to Chandragiri Mandal in the Chittoor District. The study explores whether various data mining methods are used to classify soils. The most effective strategy was contrasted with the Naive Bayes classification, as well. For soil management, agriculture, and the environment, the study's findings may be very helpful. The most effective strategy was contrasted with the Naive Bayes classification, as well. The environment, soil management, and agriculture may all benefit greatly from the study's findings.

Pusenkova et al. [2] developed the previous ten years, the yield of potatoes in Canterbury has been steady at roughly 60 t/ha. However, some commercial producers have already achieved yields more than the 90 t/ha that potato growth models predicted they would be able to produce. Over the course of two years, industry and academic partners investigated the problems restricting agricultural productivity. In year 1, 11 processing crops were closely monitored. It was found out that soil-borne diseases were a consistent source of lower yields, along with subsurface soil compaction and inefficient irrigation management. Potato fields with recent crop histories exhibited indications of Rhizoctonia stem canker appearing more quickly than those with longer crop histories. In year 2, researchers made an effort to separate and examine how soil-borne illnesses affected a commercial crop's output. . Flusulphamide, azoxystrobin, a soil fumigant, and no pesticide control were used as treatments. Results were mixed, but there was a modest reduction in Spongospora subterranea and Rhizoctonia solani DNA levels in the soil before and after treatment. The average final fresh yield per hectare was 58 t/ha and did not differ by treatment. In comparison to all previous treatments, azoxystrobin therapy reliably decreased the severity of R. solani on underground stems during the entire season.

Raymond H. Myers et al. [3] proposed collection of statistical design and numerical optimisation approaches called "response surface methodology" Plans for products and processes are optimised using (RSM). Since The process and chemical industries, in particular, have made substantial use of this research since it was initially conducted in the 1950s. RSM at this moment widely used for the past 15 years, and numerous significant breakthroughs have occurred. We concentrate on RSM efforts since 1989 in this review paper. We talk about existing research fields and suggest some areas for future study.

Dennis K. Muriithi et al. [4] examines the operational factors necessary for Kenya's highest production of potato tubers. As a result, potato farmers will gain from avoiding increasing input costs. To boost potato yield, response surface approach and factororial design were utilised. Analysis and modification of the combined impacts of water, nitrogen, and phosphorus mineral nutrients were done using response surface methods. An irrigation water level of 70.04 percent, urea-based nitrogen and triple super phosphate-based phosphorus supplies each weighing 124.75 kg per hectare each were found to be the ideal production conditions for potato tuber yield. When everything is perfect, one can produce 19.36 kg of potato tubers every 1.8 x 2.25 metre plot. In Kenya, smallholder potato farmers can increase their standard of living and avoid additional input costs by increasing their crop's production. Last but not least, This idea taken from this research on potatoes can be applied to research on other products.

Dan Li et al. [5] To find patterns in spatial yield variability, identify the main reasons why yield variability occurs, Accurate, high-resolution yield maps are essential for precision farming and offer site-specific management insights. Cultivar differences can have a significant impact when predicting potatoes' (Solanum tuberosum L.) tuber yield using remote sensing methods. This study's goal was to use machine learning techniques and cultivar information to enhance potato yield prediction using employing unmanned aerial vehicles UAVs are used for remote sensing. Various cultivars and nitrogen (N) rate testing on small plots of land were done in 2018 and 2019.

As the growing season progresses, multi-spectral photos from a UAV were gathered. Multiple vegetative metrics and cultivar characteristics were combined using machine learning models, specifically RFR (assistance vector regression) and random forest regression (SVR). It was shown that spectral information from UAV-based aircraft obtained during the beginning stages of early growth was more strongly associated with marketable output of potatoes. The optimal vegetative indexes and timing for predicting potato yield, however, differed across cultivars. When cultivar information was added, the effectiveness of the SVR and RFR models greatly increased (R2 = 0.75-0.79 for validation) compared to when only sensing data were used (R2 = 0.48-0.51). It is concluded that approaches without incorporating cultivar information perform much worse at predicting potato production compared to those that use machine learning methods to blend high spatial-resolution UAV pictures with cultivar information. More research is required to increase the accuracy of predicting potato yield.

III. METHODS

This section outlines the implementation of the planned task as well as the study's resources.

A. Dataset

This paper uses a crop prediction dataset containing 2200 records which are collected from the farming community. The dataset includes parameters such as Nitrogen, Phosphorous including environmental factors like temperature, humidity, and rainfall. Voting classifier is used for crop prediction using this dataset, which is divided with an 80 per cent to 20 per cent split, both training and testing units are divided. Informationabout the dataset, including the number of classes, class names, and dataset path, is provided in an Excel file.

B. Proposed Method

The goal of this experiment is to predict the suitable crop in the required area of agricultural land. Ensemble techniques by combining In terms of prediction accuracy, The current classification technique is outperformed by Support Vector Machine, Decision Tree, and Random Forest.

For estimating the area of cereals, kidney beans, and other energy crops that could be used to plan the layout of their planting on a farm and a national scale, the ensemble technique outperforms the present classification techniques in terms of prediction and performance.

C. Apply Algorithms

A variety of Methods from machine learning can be used on cleaned-up data, with a focus on methods that provide clear and transparent decision-making processes. Some understandable techniques include:

1. **Random Forest:** A popular ensemble option tree technique for evaluating each characteristic.
2. **Decision Tree:** Decision trees are accessible and can be depicted visually, making choosing an avenue easier.
3. **SVM (Support Vector Machine):** SVM is useful in information categorization, and its support vectors can be used to study the processes of decision-making.

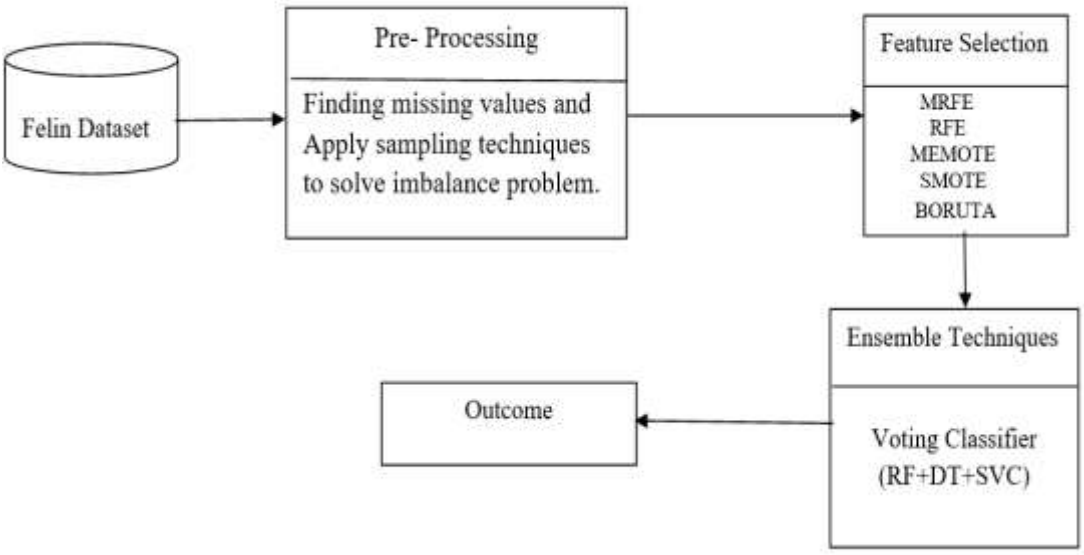

Fig. 1: Proposed System Architecture

The system architecture incorporates data collection, preprocessing, feature selection, classifier training, evaluation, and crop prediction to develop an efficient and accurate Crop Prediction system based on agricultural environment features. The system seeks to improve agricultural decision-making and crop yields by utilizing

various feature selection and ensemble technique called Voting Classifier. Felin Dataset containing 2200 records which are collected from the farming community. The dataset includes parameters such as Nitrogen, Phosphorous and environmental factors including temperature, humidity, and rainfall.

Pre- processing handles missing values and outliers in the dataset using appropriate techniques like imputation or removal. Additionally, normalizing or scaling the data is important to check all the features are on same line, which has possibility for performance improvement of some classifiers with the sampling techniques. The application of feature selection approaches allows the discovery of critical elements influencing crop forecasts. This improves the models' interpretability, allowing farmers and agricultural specialists to better comprehend the underlying causes influencing crop outcomes. Ensembling of various classifiers, allows the system to scale by the complexity of the crop prediction problem and the amount of the dataset. Because of this flexibility, advanced classifiers can be added as they become available. Voting classifier is used for crop prediction using this dataset, which is divided with an 80 per cent to 20 per cent split, both training and testing units are divided. Information about the dataset, including the number of classes, class names, and dataset path, is provided in an Excel file.

1. Random Forest (RF):

Random Forests are supervised machine learning systems that learn through decision tree approaches. It is a classification, regression, and other problem-solving ensemble learning system that functions by building many decision trees This algorithm is one of the prominent algorithm. When faced with classification challenges, the majority of trees select the Random Forest output as their class. In order to do regression tasks, the mean or average estimate of the several trees is provided. A method for reducing variation called Random Forests averages numerous trained on various subsets of the same training set, deep decision trees. This algorithm is used to predict actions and results in a range of industries, including banking and e-commerce.

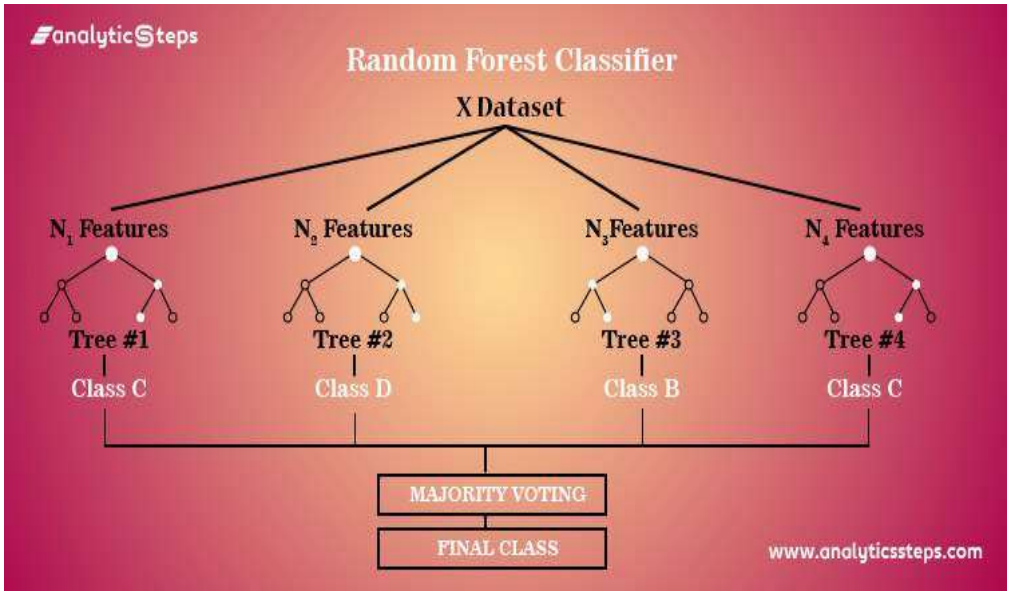

Fig. 2: Structure of Random Forest

The training examples are represented by various boxes in the Training set. These examples are used to train several Decision trees, that are shown by arrows. Decision Trees are a sort of technique that can be used for classifying and predicting data. The testing data is represented by a single box in the sample set section. This data is used to evaluate the effectiveness of the trained Decision trees. An arrow connects the field of testing data to the selecting area. During the voting process, the projections from each Decision tree are blended to form a final forecast. The Random Forest Algorithm concluded that, for classification issues the class picked by the greater number of trees. The prediction section displays the process's final output. This is where the Random Forest's ultimate result is shown. Overall, this figure illustrates the operation of a Random Forest technique. It demonstrates the use of training information to develop several Decision trees, the application of testing data in assessing their performance, and the combined effect of all of their forecasts to come up with the final forecast.

2. Decision Tree (DT)

A model called a decision tree makes predictions using a structure resembling a flowchart. It separates the data and distributes the results to the leaf nodes. Decision trees are used to develop simple models for classification and regression. The method works by repeatedly splitting the

initial information set into subsets depending on attribute values until a predetermined interruption threshold is reached, such as the top level of the hierarchy or the minimum number of occurrences necessary to divide a node. The decision tree technique determines the appropriate attribute to divide the information into segments based on a measure of quality that includes entropy or Gini impurity, which quantifies the amount of contamination or randomness in the divisions, throughout training.

Entropy(S) = $\sum_{i=1}^{c} -pi \log_2 pi$ ---------- (1)

The equation (1) shows the formula of entropy.

Gain (S, A) = Entropy(S)-$\sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(Sv)$------ (2)

The equation (2) shows the formula for information gain.

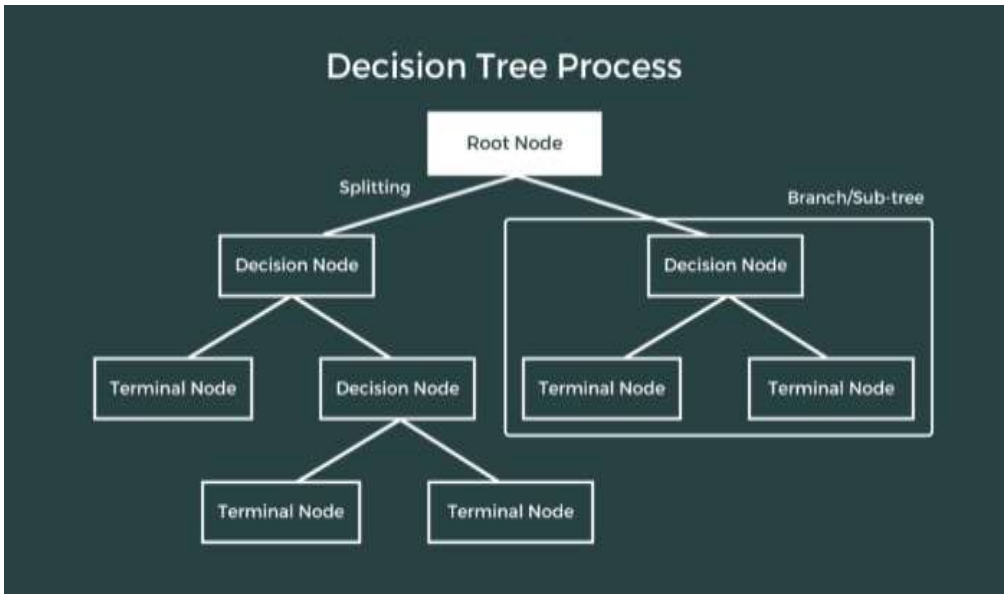

Fig. 3: Structure of Decision Tree

The Decision tree algorithm's steps are as follows:

1. S describes step one as follows: "Begin the tree at the root, which contains the entire dataset".
2. Find The dataset's most important attribute, as determined by the Attribute Selection Measure (ASM).

3. Divide groupings that could have values for the S most desirable characteristics.
4. To the node in the decision tree: Incorporate the best attribute.
5. Make a new choice. -tree structures iteratively by creating subsets of the dataset using the step 3.
6. Continue in this manner until It is no longer possible to classify the nodes and designate the final node as a leaf node.

3. Support Vector Machine (SVM)

SVM is a widely used technique. which applies for both regression and classification issues. For addressing classification issues, it is most frequently employed in machine learning. SVM is a broad topic for a calculation that works best on small but complex datasets. Support Vector Machine, sometimes known as SVM, is a technique that can be used to for planning and reversion. but it eventually proves to be too rudimentary for assembly. The SVM technique aims at finding the optimal judgement boundary or line for categorising an n-dimensional space, allowing for rapid assignment of following data points to the correct category. A hyperplane is the optimal boundary. The extreme points and vectors selected using SVM are used to build the hyperplane. . The outliers are referred to as support vectors when the Support Vector Machine technique is applied.

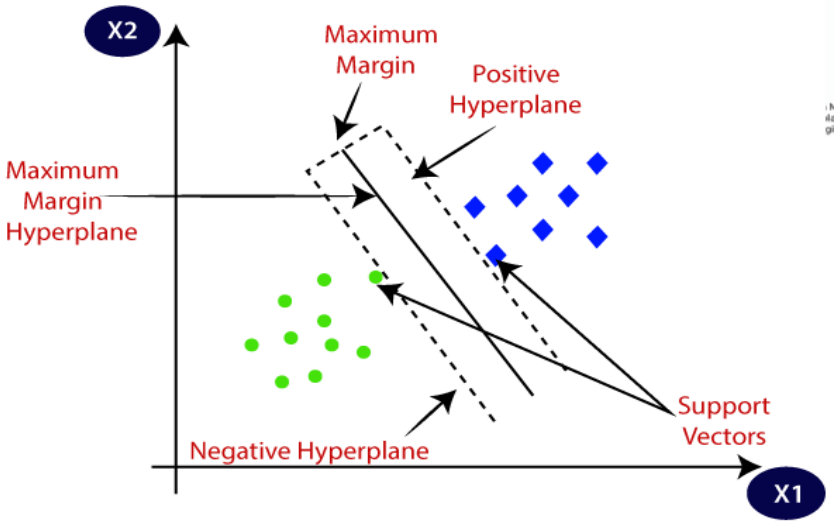

Fig. 4: support vector machine

The above figure follows the below procedure:

If the operation of an SVM classifier must be analytically understood, the following methods might be used-

1. The algorithm of the SVM predicts the classifications. The labels for one of the classes are 1, and the labels for the other are -1.

2. The business problem is transformed into a mathematical equation with unknowns, just like with earlier machine learning techniques. The unknowns are subsequently determined by approaching the topic as an optimisation problem. Since optimisation issues generally seek to maximise or minimise something when searching for and updating for unknowns, to find the highest margin, the SVM classifier updates a loss function called the loss function of the hinge.

3. When there are no classes that were mistakenly predicted, this function of loss is also known as an expense function because its cost is equal to zero. If this is not true, error/loss is computed. There is a trade-off between improving margin and the current situation, which is an issue. possibility of suffering a loss if margin is increased excessively. A regularisation parameter is provided to bring these concepts into theory.

4. Weights are optimised with other optimisation issues, by computing gradients utilising sophisticated calculus concepts such as partial derivatives. When there is no misclassification, gradients are only updated using the regularisation parameter, while those situations also involve the use of the loss function.

IV. RESULTS AND DISCUSSIONS

This paper uses a crop prediction dataset containing 2200 records which are collected from the farming community. The dataset includes parameters such as Nitrogen, Phosphorous including environmental factors like temperature, humidity, and rainfall. Voting classifier is used for crop prediction using this dataset, which is divided with an 80 per cent to 20 per cent split, both training and testing units are divided. Informationabout the dataset, including the number of classes, class names, and dataset path, is provided in an Excel file.

Table 1: Accuracy Metric Evaluation of Proposed work

S. No	Algorithm	Accuracy Achieved
1	Random Forest	99.772
2	Decision Tree	99.09
3	SVM	98.72

Table 2: Comparison with Existing work

S. No	Author	Method Used	Accuracy
1.	Raja	Random Forest	87.43
2.	Sawicka	SVM	77.50
3.	Stamenkovic	Decision Tree	73.22
4.	Mariammal	KNN	83.24
5.	Proposed Method	Voting Classifier (RF+DT+SVC)	99.772

In this work, four Evaluation Metrics were utilized to forecast crop prediction. The four measures are F1-Score, Accuracy, Precision, Recall, and Recall. Equations are used to illustrate Precision, Recall, Accuracy, and F1 Score (3-6). The work is measured using the following metrics.

$$\text{Accuracy} = \frac{TP+FP}{TP+FP+TN+FN} \quad \text{----------(3)}$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{----------(4)}$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{----------(5)}$$

$$\text{F1 Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad \text{----------(6)}$$

Where,

TP= True Positive, FP= False Positive

TN= True Negative, FN= False Negative

The experiments in this work were done using a PC with 4GB RAM, an Intel Core i55th generation CPU, and a Jupyter Notebook with 4GB storage

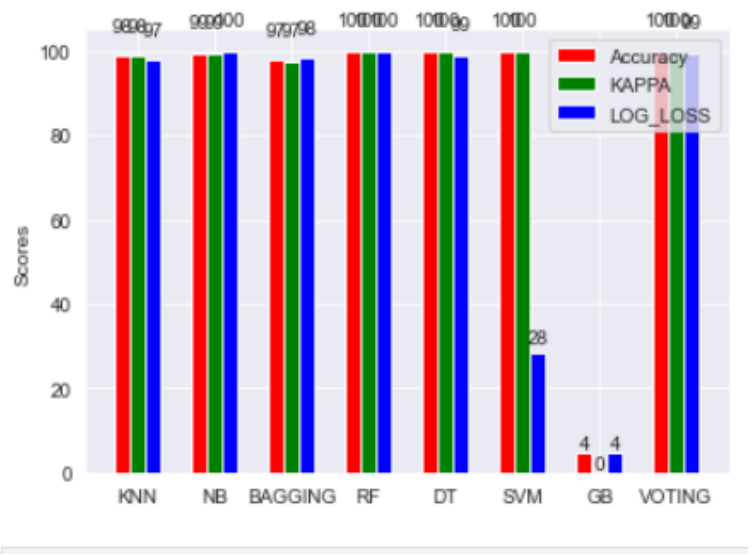

Fig. 5: Accuracy of all the algorithms

Figure. 5 Illustrates Accuracy of classification techniques and ensemble techniques, in which voting classifier has the highest accuracy of 97.7272 comparing to the other algorithms.

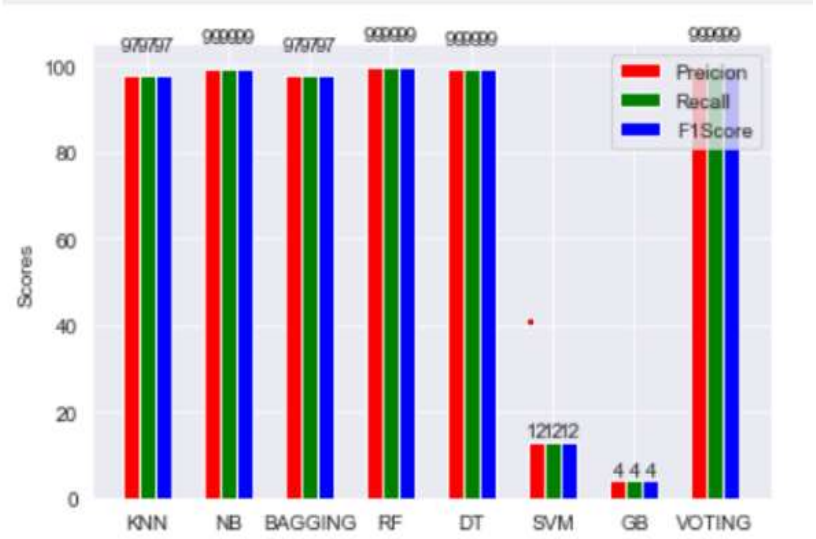

Fig. 6: Precision, recall and F1score of the algorithms

Figure 6 demonstrates that the recall and precision nd F1score of the Classification and ensemble techniques.

Based upon the above results precision, recall, and accuracy Support Vector Machine, Random Forest, and Decision Tree are all combined in the F1score Voting Classifier to provide the highest prediction rate comparing with the other

classification Techniques.

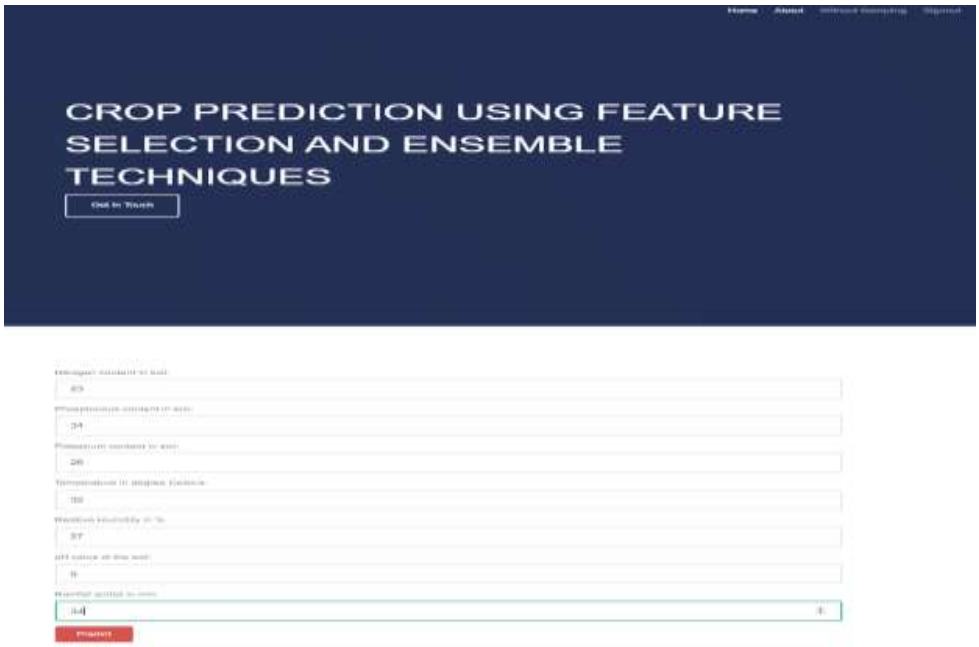

Fig. 7: Input Parameters for Predicting Crop

Figure 7 shows the input parameters which are collected to predict the crop those input parameters are Nitrogen, Phosphorous, Potassium, Temperature, Rainfall, Humidity and Ph value. Based upon all these input parameters the output will be generated after analysing the data. i.e. Summer crops & winter crops.

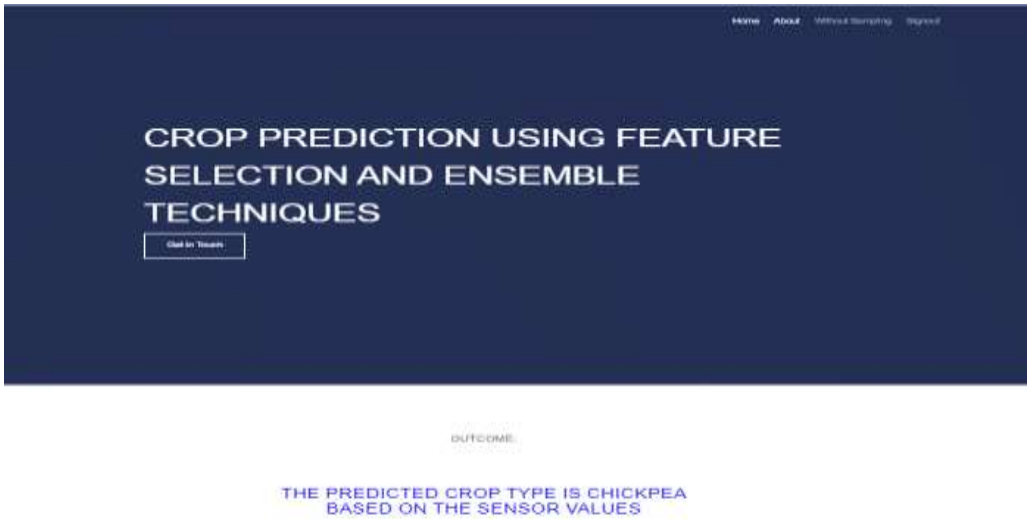

Fig. 8: Predicted Crop

Figure 8 shows the predicted crop based upon the given input parameters. It will analyze the given parameters and process the necessary steps before predicting the suitable crop.

V. CONCLUSION & FUTURE WORK

In agriculture, it can be challenging to predict which crops will grow. To determine which crop should be grown in the chosen location, a variety of feature selection and ensemble techniques have been applied. By predicting the production of potatoes, grains, and other energy crops, the sowing pattern can be planned on a farm- and a national-level basis. Utilizing modern forecasting methods can because of quantifiable monetary gains. Future research will be focused on growing the dataset's data and adding more classes in order to enhance precision, recall, and F1 score and by using sensors.

REFERENCES

[1] S.Jyothi and Peyakunta Bhargavi ‘‘Applying naive Bayes classification technique for classification of improved agricultural land soils,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. 5, pp. 189–193, May 2018.

[2] Liudmila Pusenkova, Oksana Lastochkina ‘‘Biotic components influencing the yield and quality of potato tubers,’’ Herbalism, vol. 1, no. 3, pp. 125–136, 2017.

[3] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, ‘‘Response surface methodology: A retrospective and literature survey,’’ J. Qual. Technol., vol. 36, no. 1, pp. 53–77, Jan. 2004.

[4] D. K. Muriithi, ‘‘Application of response surface methodology for optimization of potato tuber yield,’’ Amer. J. Theor. Appl. Statist., vol. 4, no. 4, pp. 300–304, 2015, doi: 10.11648/j.ajtas.20150404.20.

[5] D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, ‘‘Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning,’’ Remote Sens., vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.

[6] M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, ‘‘Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional,’’ Assoc. Agricult. Agribusiness Econ. Ann. Sci., vol. 16, no. 2, pp. 183–188, 2014.

[7] J. R. Olędzki, ‘‘The report on the state of remotesensing in

Poland in 2011–2014,'' (in Polish), Remote Sens. Environ., vol. 53, no. 2, pp. 113–174, 2015.

[8] Singh, A. K., & Ganapathysubramanian, B. (2017). Machine learning for high-throughput stress phenotyping in plants. Trends in Plant Science, 22(2), 110-124.

[9] Dash, J., & Dutta, S. (2019). A review on ensemble machine learning techniques for classification. Procedia Computer Science, 132, 377-384.

[10] Raza, S. E. A., & Antoniou, G. (2017). A survey of machine learning algorithms for big data and their applications. Journal of King Saud University-Computer and Information Sciences.

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

[12] Liu, B., & Motoda, H. (Eds.). (2008). Feature selection for knowledge discovery and data mining. Springer.

[13] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[14] Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple classifier systems (pp. 1-15). Springer.

[15] Qi, M., Qiao, F., Xiang, J., & Wu, C. (2017). A survey of crop yield prediction methods based on big data. Computers and Electronics in Agriculture, 144, 122-137.

[16] Huang, C., Davis, L. S., & Townshend, J. R. (2002). An assessment of support vector machines for land cover classification. International Journal of Remote Sensing, 23(4), 725-749.

[17] Dash, J., & Ravan, S. A. (2011). Land cover classification using remote sensing data. Remote Sensing Letters, 2(2), 147-156.

[18] Patil, P., & Mane, D. (2015). Crop yield prediction using machine learning techniques. International Journal of Computer Applications, 129(1), 32-36.

[19] Singh, A. K., & Ganapathysubramanian, B. (2017). Machine learning for high-throughput stress phenotyping in plants. Trends in Plant Science, 22(2), 110-124.