JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Best Suited Crop Recommendation SystemUsing ML

Lakshmi Shreya Bapatu Yeruguti[1], Manjunadh Konapalli[2], Manikanta Reddy Konuganti[3], Manoj Kumar Raneru[4], Suman Shekhar*, Sushil Kumar*
Department of Computer Science and Engineering
Parul Institute of Engineering and Technology, Vadodara, Gujarat, India-391760

***Abstract*—Agriculture is the mainstay of the Indian economy and employment. Unfortunately, this sector is facing significant issues due to the failure of farmers to select the appropriate crop for their soil. This has led to a drastic reduction in production. In an effort to solve this problem, precision agriculture has been used to improve crop yields and increase profitability. Precision agriculture is a modern agricultural strategy that uses research data on soil kinds, features, and crop yields to educate farmers on the best crop to grow according to specific parameters. This cutting-edge technology enables farmers to operate more efficiently by providing them with tailor-made advice and suggesting optimal solutions for their local environment. Additionally, they can also identify site-specific factors that could help them increase their crop production. By using precision agriculture, farmers can also get better insights into how to maximise their profits with limited resources. It is thus clear that precision agriculture is revolutionising the way Indian farmers operate and making it possible for them to increase their crop production significantly. As a result, crop selection errors are decreased, and production is increased. To recommend a crop for specific parameters with high accuracy and efficiency, a recommendation system based on an ensemble model with a majority voting technique might be suggested using Random Forest Tree, Decision Tree, K-Nearest Neighbor, Logistic Regression, and Naive Bayes as learners. The system's goal is to give a solution for picking a suitable crop based on weather conditions at a particular location like temperature, ph, rainfall, and soil factors including nitrogen, potassium, and phosphorous values of soil.**

***Keywords*—** recommendation, best crop, ensemble model, majority voting technique

I. INTRODUCTION

A. Definition: Suggesting the best crops for farmers to cultivate based on numerous characteristics and assist them in making a well-informed decision before production. Not taking the right decision about what to cultivate is one of the possible causes of a higher suicide rate among marginal farmers in India. They regret not getting a fruitful yield. The biggest problem facing Indian farmers is that they never choose the right crop for their soil. They are consequently dealing with a serious reduction in output. Precision agriculture has helped farmers solve many problems. Building a farmer's assistance that could suggest to farmers the type of crop that is to be sown based of the geographical location, pH values, soil type, weather patterns, etc. to get better production yield using machine learning.

B. Purpose: The purpose of the system is to provide a solution for selecting suitable crops based on the temperature, humidity, rainfall, etc. around the agricultural field. These values are given to crop recommendation assistance as input and the system determines the data and gives the results of crops to be cultivated as output. This assistance suggests the crops that have high growth through which farmers can get the maximum production and profit.

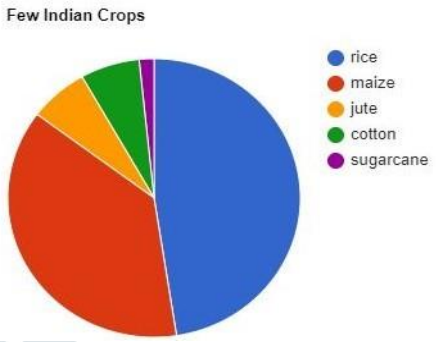

Fig. 1. **Few Major Crops**

C.Overview: The issue can be resolved by offering a recommendation system using an **ensemble model** with **majority voting technique** using a **Random forest tree**, **Decision Tree**, **K-Nearest Neighbor**, **Logistic Regression**, and **Naive Bayes**, **SVM** as learners to recommend a crop for the site specific factors with high accuracy and efficiency.

II. RELATED WORK

One method for enhancing the field of agriculture is the concept of precision farming. It's now possible to increase agricultural outputs in many places thanks to contemporary technology and methods. Using data mining to estimate agricultural productivity and assessing the effectiveness of various algorithms in the same field are only two examples of related work that has been done in the past.
The work related to crop recommendations that has been done in the past is listed below.

Title	Authors	Year	Findings
A study on various data mining techniques for crop yield prediction	*Y .Gandge and Sandhya*	2014	The SpyNB algorithm is substantially slower than an algorithm used in architecture [3].
Analysis of Soil Behaviour and Predection of Crop Yield using Data Mining Approach	*Monali Paul, Santosh K. Vishwakarma, Ashok Verma*	2015	RapidMiner is used in this study's experiments [9].
Crop predection using predective analytics	*P.S.Vijayabaskar, R.Sreemathi and E. Keertana*	2021	In accordance with the sensor's value, it also recommends the crop that should be planted. [14].
Crop recomendation system for precision agriculture system	*S. Pudumalar, E. Ramanujam, R.H.Rajashree, C. Kavya, T. Kiruthika and J. Nisha*	2017	Support Vector Machine, Random Forest, Neural Network, REPTree, Bagging, and Bayes are some of thealgorithms used. [2]
Challenges in KNN Classification	*S. Zhang*	2017	When looking for all of the K nearest neighbours for each test data, it's a complete sample space search.

III. METHODOLOGY

The suggested model makes crop production predictions by looking at variables like rainfall, temperature, nitrogen, humidity, pH values, and so forth. According to the available data for the region, it forecasts crop yield. To boost the production, proper planning and decision-making can be aided by a crop yield prediction system [3]. When predicting crop yield, the weather's impact can be seen as having a high priority. The impact of weather on agriculture has been the subject of extensive research, yet the majority of these studies call for highly complicated data that is not readily available. Due to this, data is eventually gathered via estimates [5]. Further improvements in the agriculture industry will result from the integration of ML and agriculture.

A. Project Flow:

B.

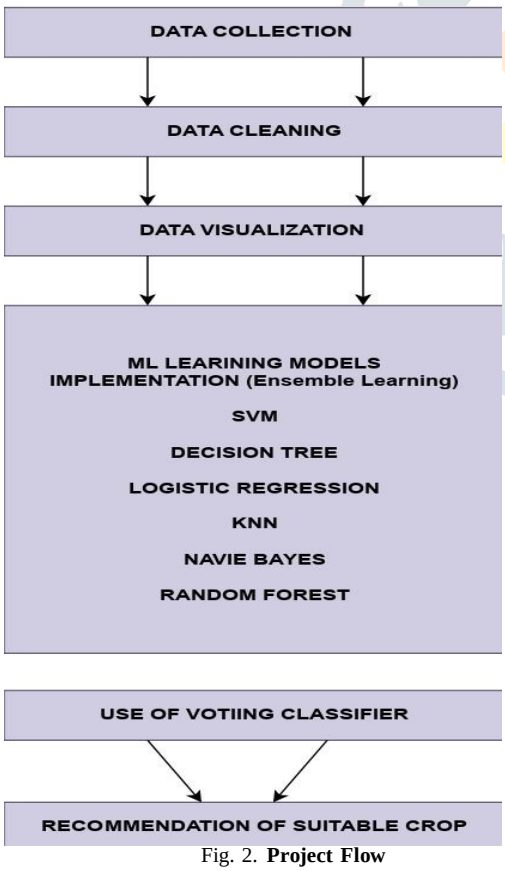

Fig. 2. **Project Flow**

Mathematical Understanding:

Two different voting methods are supported by Voting Classifier. **Hard Voting:** In a hard vote, the projected output class is the group that received the greatest number of votes or that each classifier anticipated would have the greatest chance of predicting.
In this case, the majority anticipated L as the output when three classifiers (L, L, and N) predicted the output class. Therefore, the final prediction will be L. While each classifier only votes for one class, the class receiving the most votes wins, it is occasionally referred to as a majority voting classifier. According to statistics, the mode of the individual label predictions' dispersion is the expected target label for the ensemble.

Soft Voting: In a soft vote, the forecast for the output class is based on the likelihood assigned to that class on average. Each classifier gives the likelihood that a given data item belongs to a specific target class. And the importance of the classifier is used to weight and then average the predictions. As a result, the target label with the highest importance probability and sum wins the vote.
Assume that given some input, the prediction probabilities forclasses A and B are (0.30, 0.47, and 0.53) and (0.20, 0.32, 0.40). Since class A's average is 0.4333 and class B's is 0.3067, class A isdefinitely the winner.
In a hard vote, the class that received the most votes or the class that each classifier forecasted will appear the most frequently is the projected output class.

According to this Majority Voting Classifier, the following equation represents the majority voting classifier.

$$\sum_{t=1}^{T} d_{t,J} = \max_{j=1}^{C} \sum_{t=1}^{T} d_{t,j}.$$

The classifier outputs are assumed to contain only the class labels in the discussion that follows. Defining the outcome of the t^{th} classifier as $d_{t,j} \in \{0, 1\}$, $t = 1, \ldots, T$ and $j = 1, \ldots, C$, where C represents the number of classes and T the number of classifiers. If t^{th} classifier chooses class ω_j, then $d_{t,j} = 1$, and 0 , respectively.

C. Working And Implementation:

It has been determined that Python is the optimal coding language to use for the project's system implementation. This syntax-friendly language is an excellent option for creating applications because it makes coding simpler. The decision for the project is excellent because it is currently the most well-liked programming language. Python has been demonstrated to be a dependable and efficient tool

for creating applications and machine-learning models.

Technology that is fast advancing, such as machine learning, has the potential to completely change how we interact with the outside world. As a result, learning how to use technology is now more crucial than ever in order to realize its full potential. Machine learning can be effectively used to address issues and create predictions by using Python-based modules. These libraries give programmers the tools they need to build robust algorithms and apps that have a wide range of uses. They are excellent for big projects like this one since they are also very scalable. **Machine learning** and P**ython-based Libraries** have been selected as the technology for this project because they are flexible, dependable, and potent tools.

The dataset is a crucial part of the learning models that are used to estimate the crop that would work best. Based on careful data analysis and observation, this prediction can be used to determine the best crop variety. The dataset includes details about the soil type, climate, and other factors that are important for choosing crops. By examining the data in the dataset, machine learning techniques make it possible to quickly determine the crop that is most suited for a certain region.
This dataset can also be used to create tailored recommendations for farmers depending on their particular geographic circumstances. Consequently, using this dataset to forecast the most suitable crop is crucial for agricultural research. The dataset is fed to the learning models to make a prediction of the best suited Crop.

Ensemble Learning Model is used in this project which comprises of further machine learning algorithms mentioned as follows:

a) Decision Trees
b) Random Forest Algorithm
c) Na¨ıve Bayes
d) K- Nearest Neighbour
e) Logistic Regression
f) SVM (Support Vector Machine)

It will examine the situation and recommend the most suitable crop to plant given the local climate, geography, and soil.

The Various processes involved during implementation are:

Data Cleaning: In the process of pre-processing data, data cleaning is a crucial stage. It entails locating any redundant, null, or outliers in the data and then using a variety of strategies to get rid of them. Many data-cleansing techniques are used, depending on the type of data that is accessible and the types of errors that are present in the dataset. It is crucial to remember that any kind of error in the dataset might have a significant impact on the outcomes of data analysis. Data cleaning must therefore be done before any kind of analysis.

The two stages of the data cleaning procedure are detection and removal. Potential errors are located during the detection phase utilizing a variety of statistical techniques, including descriptive statistics and visualization. Once the mistakes have been located, the

dataset's errors are removed using a variety of approaches, including imputation, discretization, normalization, encoding, and more. Data cleansing is a procedure that must be taken before performing data analysis in order to assure accurate results.

Libraries used for the same are ***Numpy*** and ***Pandas***

Two of the most popular libraries in use today are Numpy and Pandas. They have numerous uses and are an essential component of data analytics and machine learning. For scientific computing and working with big arrays and matrices, utilize Numpy. It provides a wide range of mathematical operations that facilitate the speedy resolution of challenging issues. On the other hand, Pandas are utilized for data analysis and modification. Data alignment, indexing, restructuring, merging, joining, grouping, and other aspects are among its features. Data analysis is facilitated and accelerated by the use of strong tools like Numpy and Pandas.

These libraries are used by data professionals to create predictive models and extract valuable information from the data. They can process vast volumes of data effectively without resorting to creating complicated code thanks to the combination of Numpy and Pandas. They can also deal with text processing, statistical functions, outliers, missing numbers, and much more. As a result, data analysts may complete their tasks more quickly and derive more valuable insights from their datasets.

Data Visualization: Data visualisation is an essential tool for more relevant and easily understandable data analysis and representation. It is a crucial step in the data analysis process that aids in extracting knowledge and insights from huge datasets. The libraries **Matplotlib** and **Seaborn** are frequently employed in data visualization. So far, we have created a variety of graphs using these two libraries, including **histograms**, **box plots**, **Violin plot**.

Data visualization entails presenting the data in a visual format, such as charts, graphs, maps, infographics, etc. By doing so, we are better able to spot patterns and trends in the data that might otherwise be impossible to spot when merely looking at the raw numbers. It enables us to derive inferences from the examined data and come to wise decisions. We can deliver our findings to stakeholders without difficulty thanks to data visualization, which makes it easier to communicate complex information.

Anyone who deals with data will find data visualization to be a powerful tool. It can aid in improving your understanding of your dataset and revealing intriguing patterns and connections that might not be immediately apparent. It is quite simple to build appealing visualizations that may be used for presentations or reports with the aid of programmes like ***Matplotlib*** and ***Seaborn***.

With Python and its numerical extension NumPy, there is a **cross-platform** module called **Matplotlib** that allows for **data visualisation** and **graphical plotting**. As a result, it presents a strong open-source substitute for MATLAB. The APIs (Application Programming Interfaces) for matplotlib allow programmers to incorporate graphs into GUI applications. The structure of Python code objects that makes up the plot API is headed by matplotlib.pyplot

Python has a module called ***Seaborn for creating statistical visuals.*** It is built around Matplotlib and tightly interfaces with Pandas' data structures.You may examine and comprehend your data with Seaborn.The below figure shows the histogram and boxplot of N (one of the features in the dataset). Seaborn is used for workingof one - dimensional arrays.

The below figure shows the histogram and boxplot of N (one of the features in a dataset).

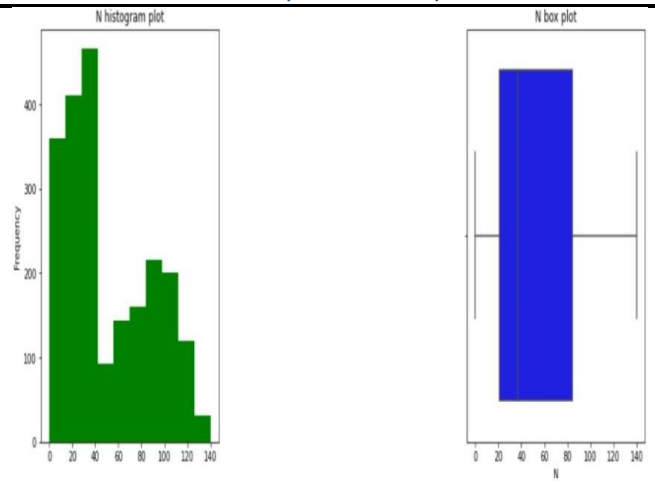

Fig. 3. **Histogram, Box plot**

The violin plot for N is given below

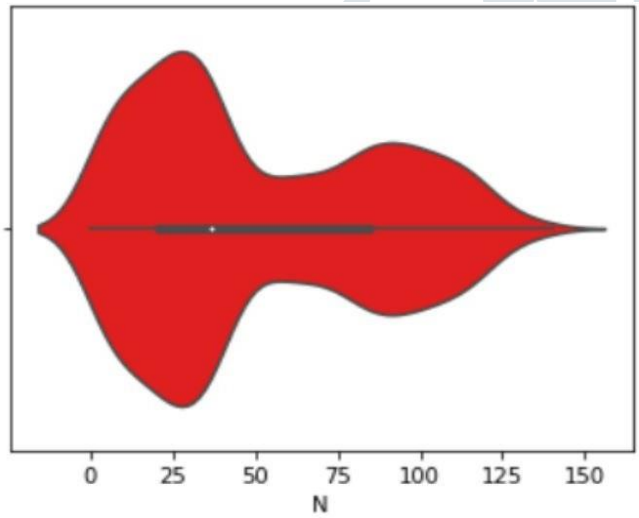

Fig. 4. **Violin Plot**

The **Voting Classifier** Ensemble Learning model is used in this system's machine learning implementation.Naive Bayes, Random Forest, SVM,Decision Trees, Logistic Regression, and KNN are some of the machine learning techniques employed in this model. The accuracy and effectiveness of the system are increased by integrating the strengths of various algorithms. The Voting Classifier's capacity to counterbalance the shortcomings of its constituent parts allows it to produce results that are superior tothose of any single model.

Using both empirical and theoretical data, the Voting The classifier approach can also be used to decide which algorithm is better suited for a certain task. Additionally, it can shed light on how several algorithms can work best when combined, allowing us to build an algorithm that outperforms its component parts. Additionally, it has been demonstrated that this method works well for enhancing the precision and effectiveness of numerous applications. Scikit Learn was utilized as the project's library. Due to its useful and simple-to-use tools, this library is frequently utilized by data scientists and machine learning specialists throughout the world. A large selection of supervised and unsupervised algorithms are included in this open source machine learning package that was created in Python.

Pre-processing, feature extraction, model selection, and evaluation are just a few of the helpful features offered by this library. Additionally, it enables users to employ feature extraction, data transformation, or feature selection either before or after estimate algorithms are used. Numerous sophisticated clustering, dimensionality reduction, and model selection methods are also included in the package.

Performing tasks like classification, regression, clustering, and anomaly detection require the use of strong tools, which Scikit Learn offers. There are several linear models available, including least squares support vector machines, logistic regression, and linear regression. Also supported are non-linear models like decision trees and kernel ridge regression.

A variety of evaluation criteria are also included in the library for gauging the effectiveness of the models. These comprise the F1 score as well as the accuracy, precision, and recall ratings. Additionally, Sci-kit Learn offers cross-validation methods that aid in evaluation

In order to create a new classifier that performs better than any of its constituent classifiers, ensemble learning generates a variety of basic classifiers.

With the decision tree and random forest algorithm, every training set is produced as a decision tree, and a random forest is subsequently created. The full decision tree predicts each sample in the testing data set [10].

Different algorithms, including ***Naive Bayes***, ***Random Forest***,***SVM***, ***Decision Trees***, ***Logistic Regression***, and ***KNN***, were put into use. To evaluate the effectiveness and accuracy of each of these algorithms, tests were run on each one separately. The majority voting classifier, which incorporates all of the aforementioned methods, was our final choice. This approach offers a more precise prediction since it considers the consensus of all algorithms rather than just one. As it uses numerous models to obtain its outcome, the majority voting classifier also reduces overfitting. By using the majority vote from several models that are trained on various subsets of data, it aids in class labeling.

This method is advantageous for lowering the bias and variance linked to specific algorithms.

We can get a better prediction by using the majority voting classifier since it considers the opinions of all the algorithms being utilized as a whole.

The Majority Voting Classifier has been a crucial part of our model's success. We have found that this approach has significantly improved our accuracy and allowed us to reach results that were unattainable using any single algorithm. Additionally, it has helped us create a robust system that can withstand changes in the data distribution and work reliably without requiring any manualinterventions.

D. Comparison Of Algorithms:

The algorithms implemented and their respective accuracy is givenin the below table.

Algorithm	Accuracy	Input Data	Output
SVM	10.68%	[90,42,43, 23.603016,60.3, 6.7,140.91]	Kidney Beans
Logistic Regression	95.22%	[90,42,43, 23.603016,60.3, 6.7,140.91]	Jute
Random Forest	98.63%	[90,42,43, 23.603016,60.3, 6.7,140.9]	Coffee
K- Nearest Neighbour	95.45%	[90,42,43, 23.603016,60.3, 6.7,140.91]	Jute
Na¨ıve Bayes	90.45%	[90,42,43, 23.603016,60.3, 6.7,140.91]	Jute
Decision Tree	90.45%	[90,42,43, 23.603016,60.3, 6.7,140.91]	Jute
Ensemble Technique	**98.86%**	**[90,42,43, 23.603016,60.3, 6.7,410.91]**	**Jute**

Highest accuracy among the above independent machine learning algorithms were found to be 98.63% **(Random Forest)** and the the lowest accuracy was found to be approximately 10.68% **(SVM)**. But, to improve the accuracy by some more margin we have used Voting Classifier which takes all the ml algorithms into consideration and opts out of a result that is best for the given input values.

The Final Result: The Accuracy for Voting Classifier model is 98.86% and JUTE is the corresponding output recommendation for the given input data points of N, P, K, temperature, pHvalue, rainfall

Decision trees are easily understood and explained. purposes. Unlike most other classification systems, decision trees make it simple for users to prepare their data algorithms [6]. The decision tree is made up of nodes, with the root node serving as a representation for all the rows in a dataset. The next step involves utilising a splitting variable, also known as child nodes, to divide each node into two nodes. Recursive partitioning is being used here. Terminal or leaf nodes are nodes that do not have child nodes. The target variable's values are carried by leaf nodes. The fundamental benefit of DT models is that they can aid non-technical persons in comprehending the scope of a specific issue. The main drawback of DT models is that every single node is only individually optimised whereas the entire tree is globally optimised [7]. A small decision tree with just the right number of rules to be able to classify data and make predictions with some degree of accuracy and interpretability [16].

KNN was selected because it performs well with tiny dataset [4]. A few challenges in individual machine algorithms used above are that in KNN generally, there are four difficult problems: (i) The categorization rule, (ii) K computation, (iii) nearest neighbour search, (iv) and nearest neighbour selection. [18]. Given that it is challenging to calculate estimates of probability densities, KNN classification provides discriminant analysis. These are the identical data points that were used for training in the previous case. New, unlabeled data are offered for testing purposes. Here, the goal is to determine the new point's class label. According to the value of k, diverse results are obtained [12]. Recently, several attempts have been made to set the K value or find K nearest neighbours in order to address the lazy aspect of KNN classification [20].

The **Random Forest** ensemble approach, which combines the output of numerous decision trees, is used to produce a forecast. In a Random Forest, there are two parameters: L, which establishes the ensemble's size, and attempts,which denotes the number of distinct characteristics that are randomly chosen at each node.An algorithm is used to build each tree in the forest [19]..While RFS (Random Forest) has demonstrated remarkable practical performance in a variety of real-world applications, nothing is known about its theoretical features. Consistency is the most important theoretical requirement for a learning algorithm since it ensures convergence to the best solution as the data grows indefinitely vast. It is difficult to demonstrate the consistency of RFs because they use randomized instance bootstrapping, randomized feature bagging, and deterministic tree construction [15].

For **SVM**, It becomes difficult to imagine when the number of features exceeds three, and in our case, the feature is six.Using data from previous events, a method known as naive Bayes probability calculates the probability of future events [17]. For a number of uses in data mining and machine learning, the Naive Bayes classifier (NBC) is utilised extensively. It has been shown to be astonishingly durable. Its performance has been somewhat puzzling due to this given its strong independence assumption on the characteristics [11].

The assumption of predictor independence is one of the primary causes behind the **Naive Bayes** Classifier's superior performance. The Naive Bayes Classifier sometimes loses accuracy because of this very independent assumption. When data sets contain attributes that strongly relate to one another, accuracy loss may be greater. As a result, it is difficult to increase accuracy in a Naive Bayes classifier under the assumption that predictors are independent [8].

Logistic regression falls under the topic of **Supervised Learning** and is one of the most well-known Machine Learning algorithms. It is used to forecast the categorical dependent variable using a specified set of independent variables. In the output, a categorical dependent variable, is forecast using logistic regression. The outcome must therefore be a discrete or categorical value. It offers the probabilistic values that lie between 0 and 1 rather than the precise values between0 and 1. It can be either **True** or **False**, 0 or 1, or Yes or No.

Every machine algorithm has a unique set of benefits and drawbacks. Hence, by using the majority voting classifier we aim to select the result from a particular algorithm that suits the best-given input.

The voting classifier includes the models SVM, Logistic Regression, Random Forest, K-Nearest Neighbour, Naive Bayes, and Decision Tree models to produce the prediction that is most accurate and appropriate for input.

The voting classifier is a particular kind of machine learning estimate that creates several base models or estimation method and then generates predictions by averaging their outcomes. The generalization capabilities of learning classifier systems have been found to be greatly enhanced by the use of a rough set attribute-reduction-based ensemble technique. This method is perfect for handling high-dimensional datasets where the sheer number of characteristics might be daunting. A basic set attribute-reduction-based ensemble technique can be used to minimize the feature space, allowing for greater generalization and more accurate classification. [1].

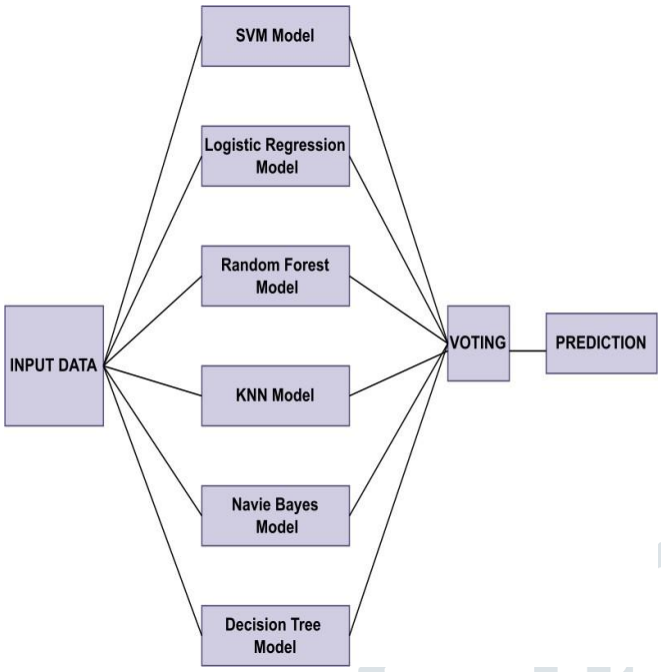

Fig. 5. **Voting Classifier**

To implement the voting classifier for our system we need to import it first from sckit-learn and then the below mentioned code is to be used such that all the individual models (Support Vector Machine, Na¨ıve Bayes, Decision tree, Random Forest, KNN, and Logistic Regression) is given to the ensemble model. The ensemble model determines the outcome of the most accurate model and provides the final choice based on the input data.

Instead of creating individual specialised models and evaluating their individual correctness, the goal is to create a single model that draws knowledge from other models and predicts results based on the sum of their votes for each output class. When there are few positive cases and a significant number of negative instances, often known as imbalanced data sets, the ensemble approach generally performs reasonably well [13]. Since voting classifier(ensemble model) has given us the highest accuracy among all other individual algorithms we take its output as the consideration for the most suitable crop that could be grown at a place by the farmer.

IV. CONCLUSION

This project proposes to recommend the type of crop to be grown at a certain place (Precision Agriculture) based on several conditions such as rainfall, temperature, nitrogen, potassium, phosphorous, and pH values, etc. we are **using Ensemble Learning model- Voting Classifier**. In conclusion, using a voting classifier offers an effective solution to address complicated issues and raises the system's accuracy and effectiveness. It is a useful technique for getting outcomes that are superior to what any algorithm might have obtained on its own. As a result, it is strongly advised for more effective implementation of machine learning tasks. The use of precision agriculture has already resulted in a notable revolution in the agricultural sector. Through the use of sophisticated tools for forecasting and decision-making, farmers are now able to maximize their output and cut expenditures.

V. FUTURE SCOPE

Precision agriculture has become better and has a larger edge over traditional methods thanks to the use of deep learning algorithms like **Artificial Neural Networks** (ANN) and **Multi-Layer Perceptrons** (MLPs) have improved precision agriculture and given it a stronger advantage over conventional techniques.Thanks to technological advancements, deep learning algorithms can now be used extensively in precision agriculture. To guarantee that the algorithms' predictions are accurate, the data must be collected, cleaned, and properly labeled. In addition, precise data collection from the environment will require high-quality sensors. Deep Learning algorithms are a promising technique for precision farming, offering several benefits like increased efficiency and accuracy. And this remains as our future scope with respect to this project.

REFERENCES

[1] E. Debie, K. Shafi, C. Lokan, and K. Merrick. Reduct based ensemble of learning classifier system for real-valued classification problems. In *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*, pages 66–73. IEEE, 2013.

[2] A. S. Dhevi. Imputing missing values using inverse distance weighted interpolation for time series data. In *2014 Sixth international conference on advanced computing (ICoAC)*, pages 255–259. IEEE, 2014.

[3] Y. Gandge and Sandhya. A study on various data mining techniques for crop yield prediction. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 420–423. IEEE, 2017.

[4] N. Guntamukkala, R. Dara, and G. Grewal. A machine-learning based approach for measuring the completeness of online privacy policies. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*, pages 289–294. IEEE, 2015.

[5] R. Gupta, A. K. Sharma, O. Garg, K. Modi, S. Kasim, Z. Baharum, H. Mahdin, and S. A. Mostafa. Wb-cpi: Weather based crop prediction in india using big data analytics. *IEEE Access*, 9:137869–137885, 2021.

[6] K. Lavanya, S. Bajaj, P. Tank, and S. Jain. Handwritten digit recognition using hoeffding tree, decision tree and random forests—a comparative approach. In *2017 international conference on computational intelligence in data science (ICCIDS)*, pages 1–6. IEEE, 2017.

[7] A. B. Nassif, M. Azzeh, L. F. Capretz, and D. Ho. A comparison between decision trees and decision tree forest models for software development effort estimation. In *2013 Third International Conference on Communications and Information Technology (ICCIT)*, pages 220–224. IEEE, 2013.

[8] K. Netti and Y. Radhika. A novel method for minimizing loss of accuracy in naive bayes classifier. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–4. IEEE, 2015.

[9] M. Paul, S. K. Vishwakarma, and A. Verma. Analysis of soil behaviour and prediction of crop yield using data mining approach. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 766–771. IEEE, 2015.

[10] S. Sahu, M. Chawla, and N. Khare. An efficient analysis of crop yield prediction using hadoop framework based on random forest approach. In *2017 international conference on computing, communication and automation (ICCCA)*, pages 53–57. IEEE, 2017.

[11]C. R. Stephens, H. F. Huerta, and A. R. Linares. Why the naive bayes approximation is not as naive as it appears. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–6. IEEE, 2015.

[12] K. Swati and A. Patankar. Effective personalized mobile search using knn. In *2014 international conference on data science & engineering (ICDSE)*, pages 157–160. IEEE, 2014.

[13] C. K. Veni and T. S. Rani. Ensemble based classification using small training sets: A novel approach. In *2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL)*, pages 1–8. IEEE, 2014.

[14] P. Vijayabaskar, R. Sreemathi, and E. Keertanaa. Crop prediction using predictive analytics. In *2017 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC)*, pages 370–373. IEEE, 2017.

[15] Y. Wang, S.-T. Xia, Q. Tang, J. Wu, and X. Zhu. A novel consistent random forest framework: Bernoulli random forests. *IEEE transactions on neural networks and learning systems*, 29(8):3510–3523, 2017.

[16] H. Yang and S. Fong. Optimized very fast decision tree with balanced classification accuracy and compact tree size. In *The 3rd international*

conference on data mining and intelligent information technology applications, pages 57–64. IEEE, 2011.

[17] B. W. Yohanes, S. Y. Rusli, and H. K. Wardana. Location prediction model using na¨ıve bayes algorithm in a half-open building. In *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pages 15–19. IEEE, 2017.

[18] J. Zhang, H. Qi, Y. Ji, Y. Ren, M. He, M. Su, and X. Cai. Nonlinear acoustic tomography for measuring the temperature and velocity fields by using the covariance matrix adaptation evolution strategy algorithm. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14, 2021.

[19] L. Zhang, Y. Ren, and P. N. Suganthan. Instance based random forest with rotated feature space. In *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*, pages 31–35. IEEE, 2013.

[20] S. Zhang and J. Li. Knn classification with one-step computation. *IEEE Transactions on Knowledge and Data Engineering*, 2021.