BEST SUITED CROP RECOMMENDATION SYSTEM USING ML

A Project Report

Submitted By

BAPATU YERUGUTI LAKSHMI SHREYA
KONAPALLI MANJUNADH
KONUGANTI MANIKANTA REDDY
RANERU MANOJ KUMAR

in Partial Fulfilment For the Award of

the Degree of

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE & ENGINEERING

Under the Guidance of

Prof.(Mr.) Sushil Kumar

Associate Professor

PARUL UNIVERSITY

VADODARA

February - 2023

PARUL UNIVERSITY

CERTIFICATE

This is to Certify that Project - 2 -Subject code 203105400 of 8th Semester entitled “Best Suited Crop Recommendation System Using ML” of Group No. PUCSE_30 has been successfully completed by

- BAPATU YERUGUTI LAKSHMI SHREYA- 190304105069
- KONAPALLI MANJUNADH- 190304105084
- KONUGANTI MANIKANTA REDDY- 190304105085
- RANERU MANOJ KUMAR- 190304105097

under my guidance in partial fulfillment of the Bachelor of Technology (B.TECH) in Computer Science and Engineering of Parul University in Academic Year 2022- 2023.

Date of Submission :-

Mr Sushil Kumar,

Project Guide

Associate Professor

Project Coordinator:-

Head of Department,

Dr. Amit Barve

CSE, PIET

Parul University

External Examiner

ACKNOWLEDGEMENT

Behind any major work undertaken by an individual there lies the contribution of the people who helped to cross all the hurdles to achieve the goal. It gives us immense pleasure to express my sense of sincere gratitude towards our respected guide Mr. Sushil Kumar, Assistant Professor for his persistent, outstanding, invaluable cooperation and guidance. It is our achievement to be guided under him. He is a constant source of encouragement and momentum that any intricacy becomes simple. We gained a lot of invaluable guidance and prompt suggestions from him during the entire project work. We will be indebted of him forever and we take pride to work under him.

We also express our deep sense of regards and thanks to Mr.Amit Barve (Professor) and Head of CSE Engineering Department. We feel very privileged to have had their precious advice, guidance, and leadership.

PLACE: VADODARA

DATE:

BAPATU YERUGUTI LAKSHMI SHREYA - 190304105069

KONAPALLI MANJUNADH - 190304105084

KONUGANTI MANIKANTA REDDY - 190304105085

RANERU MANOJ KUMAR - 190304105097

Contents

List of Figures

List of Tables

Chapter 1

ABSTRACT

Agriculture in India plays a predominant role in economy and employment. The common problem existing among the Indian farmers are they don't choose the right crop based on their soil requirements. Due to this they face a serious setback in productivity. This problem of the farmers has been addressed through precision agriculture. Precision agriculture is a modern farming technique that uses research data of soil characteristics, soil types, crop yield data collection and suggests the farmers the right crop based on their site- specific parameters. This reduces the wrong choice on a crop and increase in productivity. This problem can be solved by proposing a recommendation system through an ensemble model with majority voting technique using Random Forest tree, Decision Tree, K-Nearest Neighbor, Logistic Regression and Naive Bayes as learners to recommend a crop for the site-specific parameters with high accuracy and efficiency.

Chapter 2

Introduction

2.1 Definition

To recommend the Best suited crops to be cultivated by farmers based on several parameters and help them make an informed decision before cultivation. Not taking the right decision of what to cultivate is one of the possible causes for a higher suicide rate among marginal farmers in India. They regret after not getting a fruitful yield. The common problem existing among the Indian farmers are they don't choose the right crop based on their soil requirements. Due to this they face a serious setback in productivity. This problem of the farmers has been addressed through precision agriculture. Building a farmer's assistance that could suggest farmers the type of crop that is to be shown based of the geographical location, weather patterns, pH values, soil type etc. to get better production yield using machine learning.

2.2 Purpose

The purpose of the system is to provide the solution for selecting the suitable crop based on the Temperature, Humidity, Rainfall etc. around the 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 has high probability of growth through which farmers can get the maximum production and profit.

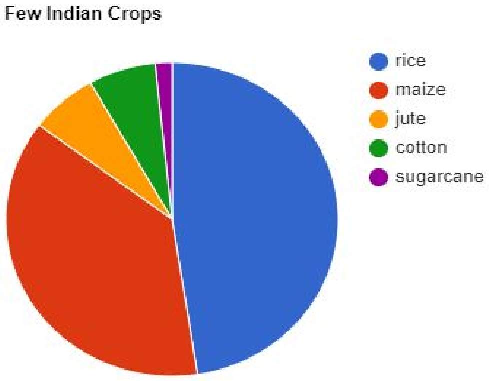


Figure 2.1: Few Major Crops.

2.3 Scope

• The Scope of this project is, that it will be used to take better decisions over taking random decisions on type of crop to be cultivated by farmers.

• The main objective is to obtain a better variety of crops that can be grown over the season.

• The proposed system would help to minimize the difficulties faced by farmers in choosing a crop and maximize the yield.

2.4 Overview

The proposed model provides crop selection based on environmental conditions, and benefit to maximize the crop yield that will subsequently help to meet the increasing demand for the country's agricultural needs. The proposed model predicts the crop yield by studying factors such as rainfall, temperature, area, season, soil type, humidity, pH values etc. It predicts the crop yield for the data sets of the given region. Integrating agriculture and ML will contribute to more enhancements in the agriculture sector by increasing the yields and optimizing the resources involved. The data from previous years are the key elements in prediction of current performance.

Chapter 3

LITERATURE REVIEW

1. Optimized very fast decision tree with balanced classification accuracy and compact tree size

   Authors: H. Yang and S. Fong

   - Optimized-VFDT algorithm that uses an adaptive tie mechanism to automatically search for an optimized amount of tree node splitting, balancing the accuracy and the tree size, during the tree-building process.
   - The aim of optimizing the algorithm is to achieve an optimal balance between accuracy and tree size.

2. A Novel Consistent Random Forest Framework: Bernoulli Random Forests

   Authors: Y. Wang, S. -T. Xia, Q. Tang, J. Wu and X. Zhu

   - The key factor lies in the two Bernoulli driven/controlled tree construction processes. A certain degree of randomness as well as the overall quality of the trees is ensured simultaneously.
   - The training data set D is first partitioned into a Structure part and an Estimation part for achieving the consistency property of

the proposed BRF.

3. Why the Naive Bayes approximation is not as Naive as it appears

   Authors: C. R. Stephens, H. F. Huerta and A. R. Linares

   - Analyze under what circumstances the NBA and associated NBC can be expected to be suboptimal and develop general diagnostics with which a problem can be examined.
   - It also allows us to clearly see that attribute correlation does not necessarily lead to inferior performance of the NBA due to the possibility of error cancelation.

4. An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach

   Authors: S. Sahu, M. Chawla and N. Khare

   - There remains a need for distance function for the selection of the K nearest neighbor points that can work effectively across most training samples.
   - It is a complete sample space search when looking for all the K nearest neighbors for each test data. As a result, KNN classification is often referred to as a lazy data mining method.

5. Crop prediction using predictive analytics

   Authors: P. S. Vijayabaskar, R. Sreemathi and E. Keertanaa

- Predictive model often perform calculation during the transaction in order to evaluate the risk and opportunity of customer during the transaction.
- Using classification, multiple linear regressions, prediction, decision tree algorithm prediction of the agriculture was done.

6. A comparison between decision trees and decision tree forest models for software development effort estimation
   Authors: B. Nassif, M. Azzeh, L. F. Capretz and D. Ho

   - DT models might suffer from the overfitting problem, as well as from providing good accuracy in comparison to other models.
   - A DTF model is similar to a Tree boost model in the sense that both models use a large number of trees

7. WB-CPI: Weather Based Crop Prediction in India Using Big Data Analytics
   Authors: R. Gupta et al

   - The main aim would be to process the data using MapReduce and frame a recommender algorithm in Python to extract output.
   - MapReduce model was implemented on the cleaned data, where the dataset is divided into key and column pairs. attribute reduction presents systematic method to get set of attribute subsets that do not lose distinguishing information in the original data.

8. A Machine-Learning Based Approach for Measuring the Completeness of Online Privacy Policies.

   Authors: N. Guntamukkala, R. Dara and G. Grewal

   - Classifiers were chosen to perform the text classification task for completeness. The size of the dataset, and the number of features were all taken into consideration
   - The accuracy results suggest that LSVM and KNN outperform RF.

9. K-nearest neighbor-based bagging SVM pruning

   Authors: R. Ye, Z. Le and P. N. Suganthan

   - The lazy bagging approach is compared with conventional bagging SVM and the results show a positive impact of lazy bagging.
   - Automated approach to evaluate the completeness of privacy policies to empower users to take greater control of their data.

10. Challenges in KNN Classification

    Authors: Y. S. Zhang

    - There remains a need for distance function for the selection of the K nearest neighbor points that can work effectively across most training samples.
    - It is a complete sample space search when looking for all the K nearest neighbors for each test data. As a result, KNN classification is often referred to as a lazy data mining method.

11. Crop recommendation system for precision agriculture

    Authors: S. Pudumalar,E. Ramanujam, R. H. Rajashree,C. Kavya,T. Kiruthika and J. Nisha

    - The algorithms used for yield prediction in this paper are Support Vector Machine, Random Forest, Neural Network, REPTree, Bagging, and Bayes
    - The rules are generated in form of if-then rules where the then part specifies the class label. The rules generated from the above model is used to develop a RECOMMENDATION SYSTEM

12. Instance based random forest with rotated feature space

    Authors: L. Zhang, Y. Ren and P. N. Suganthan

    - The main idea is to improve the diversity among the weak classifiers in the ensemble and eliminate several weak classifiers which are not the most appropriate to the particular test sample form the ensemble.
    - The rotated features are used to generate the Random Forest model and the decision trees are really sensitive to the rotation, this rotation approach will improve the diversity among the trees in the forest.

13. Reduct based ensemble of learning classifier system for real-valued classified problem

    Authors: E. Debie, K. Shafi, C. Lokan and K. Merrick

- Rough set attribute reduction-based ensemble approach used into learning classifier system to improve generalization capabilities.
- Rough set attribute reduction presents systematic method to get set of attribute subsets that do not lose distinguishing information in the original data.

14. Ensemble based classification using small training sets: A novel approach
    Authors: C. V. Krishna Veni and T. S. Rani

    - Generating a core set of instances or representative set that can be used to train a classifier with very small number of instances without losing generality is the main goal of this paper.
    - If we are able to choose representative samples even 10% to 18% of the data set can train a classifier efficiently

15. Handwritten digit recognition using hoeffding tree, decision tree and random forests — A comparative approach
    Authors: K. Lavanya, S. Bajaj, P. Tank and S. Jain

    - Decision trees provide one of the simplest portrayals for the classification purposes. The dataset initially presented is split into smaller units alongside the creation of a correlated decision tree in cumulative manner.
    - After the building of every tree, the data is passed to the entire tree and closeness, also known as proximity, is computed for every single pair of case.

16. A novel method for minimizing loss of accuracy in Naive Bayes classifier
    Authors: K. Netti and Y. Radhika

    - Conditional Independence is the one of the reasons for Loss of Accuracy in Naïve Bayes Classifier.
    - The conclusion, based on the experimental results is that, accuracy of Naïve Bayes classifier can be improved even with the assumption of Conditional Independence.

17. Location prediction model using Naïve Bayes algorithm in a half-open building
    Authors: B. W. Yohanes, S. Y. Rusli and H. K. Wardana

    - Naïve Bayes is used to find the location of the laptop when the RSS data is obtained. Offline phase or training phase is used to collect RSS values .
    - The smaller number of AP give a bad predicted result; therefore, a greater number of AP also need to be considered in the process of location prediction.

18. KNN Classification with One-step Computation
    Authors: S. Zhang and J. Li

    - The nearest neighbor search, researchers have done a lot of work in this area, and they improve KNN by proposing new distance measurement functions.
    - Introduce the proposed objective function to obtain the optimal K value of each test data.

19. Effective personalized mobile search using KNN

    Authors: K. Swati and A. J. Patankar

    - To predict the preferences of the user from clickthrough data for personalized query obtained from users
    - An algorithm used in architecture is much faster than the SpyNB algorithm

20. A study on various data mining techniques for crop yield prediction

    Authors: Y. Gandge and Sandhya

    - To predict the preferences of the user from clickthrough data for personalized query obtained from users.
    - An algorithm used in architecture is much faster than the SpyNB algorithm

Title	Authors	Year	Findings
A study on various data mining techniques for crop yield prediction.	*Y.Gandge and Sandhya*	2014	An algorithm used in architecture is much faster than the SpyNB algorithm [1].
Analysis of soil Behaviour and Prediction of crop Yield using Data Mining Approach	*Monali Paul Santosh K. Vishwakarma, Ashok Verma*	2015	In this work the experiments are performed using RapidMiner[3].
Crop Prediction using predictive analytics	*P.S.Vijayabaskar , R.Sreemathi and E.Keetana*	2021	It also suggests the crop which has to be planted depending upon the value obtained from the sensor[4].
Crop recommendation system for precision agriculture	*S.Pudumalar, E.Ramanujam, R.H.Rajashree, C. Kavya, T.Kiruthika and J.Nisha*	2017	The algorithms used are Support Vector Machine, Random Forest,neural Network, REPTree, Bagging,and Bayes[5]
Challenges in KNN Classification	*S.Zhang*	2017	It is a complete sample space search when looking for all the K nearest neighbour for each test data.

Table 3.1: Related Work

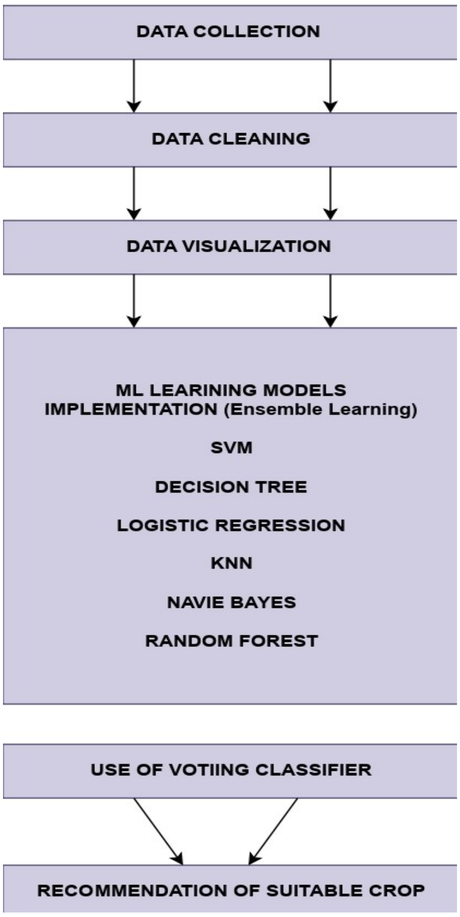
Chapter 4

PROJECT FLOW AND METHODOLOGY

The proposed model predicts the crop yield by studying factors such as rainfall, temperature, nitrogen, humidity, pH values etc. It predicts the crop yield for the data of the given region. A crop yield prediction system can help with better planning and decision-making to increase the yield.[1]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[16].The Integration of agriculture and ML will contribute to more enhancements in the agriculture sector.

4.1 Project Flow:

.


DATA COLLECTION
DATA CLEANING
DATA VISUALIZATION
ML LEARINING MODELS IMPLEMENTATION (Ensemble Learning)
SVM
DECISION TREE
LOGISTIC REGRESSION
KNN
NAVIE BAYES
RANDOM FOREST
USE OF VOTIING CLASSIFIER
RECOMMENDATION OF SUITABLE CROP

4.2 Mathematical Understanding:

Two different voting methods are supported by Voting Classifier Hard Voting: A class with the highest majority of votes, or the class that had the highest likelihood of being predicted by each of the classifiers is the projected output class in a hard vote.

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.It is also known as majority voting classifier every individual classifier votes for a class, and the majority wins. In statistical terms, the predicted target label of the ensemble is the mode of the distribution of individually predicted labels.

Soft Voting: In a soft vote, the forecast for the output class is based on the likelihood assigned to that class on average. Every individual classifier provides a probability value that a specific data point belongs to a particular target class. The predictions are weighted by the classifier's importance and summed up. Then the target label with the greatest sum of weighted probabilities wins the vote.

Assume that given some input, the prediction probabilities for classes 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 is definitely the winner.

A class with the highest majority of votes, or the class that had the highest likelihood of being predicted by each of the classifiers, is the projected output class in a hard vote.

In this Majority Voting Classifier: The equation of majority voting classifier is given below.

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

In the following discussion, we assume that only the class labels are available from the classifier outputs. Let us define the decision of the t^{th} classifier as $d_{t,j} \in 0, 1$, $t = 1 \ldots, T$ and $j = 1, \ldots C$, where T is the number of classifiers and C is the number of classes. If t^{th} classifier chooses class ω_j, then $d_{t,j} = 1$, and 0 , otherwise.

4.3 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 Python-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) Naïve Bayes

d) K- Nearest Neighbour

e) Logistic Regression

f) SVM (Support Vector Machine)

It will analyze and recommend the best crop to be grown with certain environmental, geographical, and soil conditions.

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. Depending on the type of data available and the type of mistakes in the dataset, several data-cleaning approaches are employed. 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: In order to analyze and depict data in a way that is more meaningful and simple to comprehend, data visualization is a crucial tool. 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.

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open-source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications. The plot API is a hierarchy of Python code objects topped by matplotlib.pyplot

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data. The below figure shows the histogram and boxplot of N (one of the features in the dataset). Seaborn is used for working of one - dimensional arrays.

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

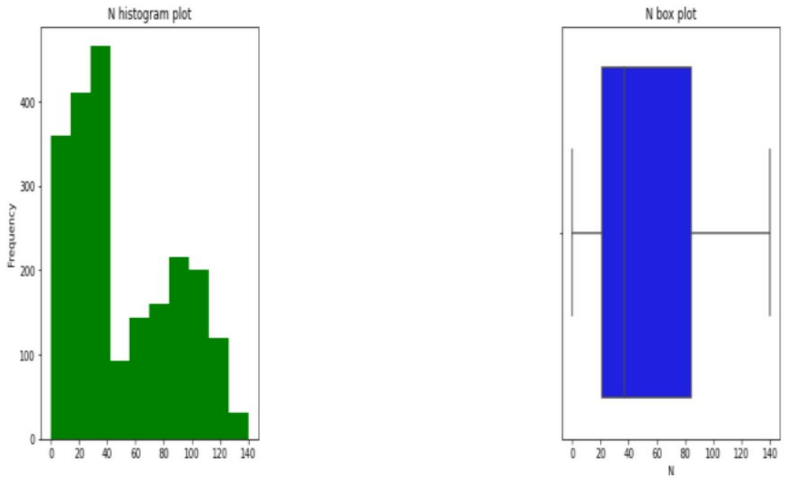


Figure 4.1: Histogram

The violin plot for N is given below

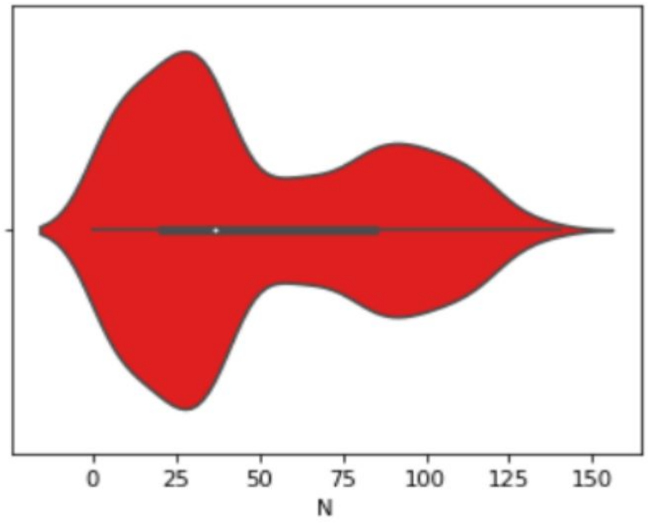


Figure 4.2: Violin Plot

The Voting Classifier Ensemble Learning model is used in this system's machine learning implementation.The machine learning algorithms used in this model are Naive Bayes, Random Forest, SVM, Decision Trees, Logistic Regression, and KNN. 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 to those 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.

In the decision tree and random forest, each and individual training set is constructed as the decision tree, finally, a random forest is generated. Each sample of the testing data set is predicted by the entire decision tree.[2]

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 manual interventions.

4.4 Comparison Of Algorithms:

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

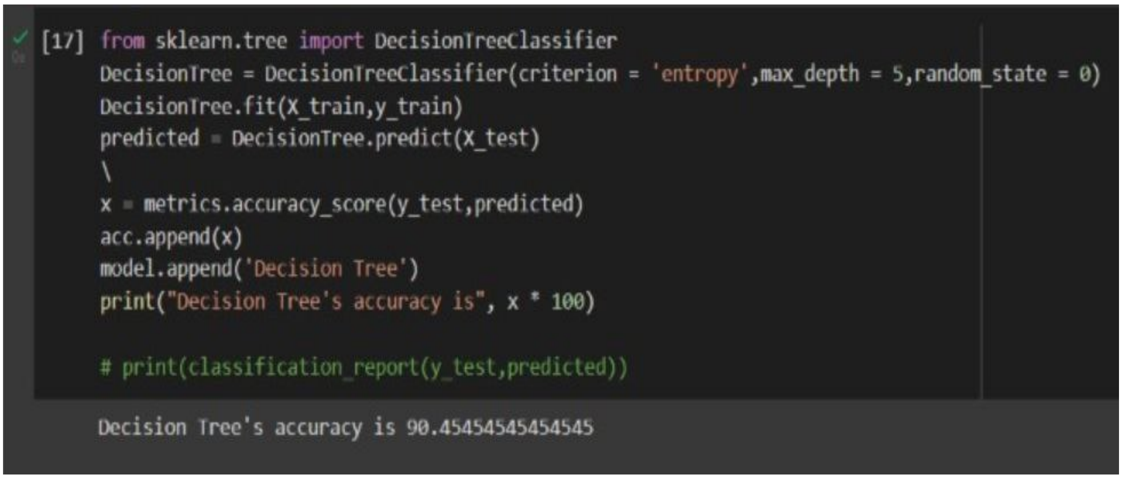
Algorithm	Accuracy	Input Data	Output
SVM	10.68%	[90,42,43, 23.603016, 60.3, 6.7, 140.91]	Kidney Beans
Logistic Regression	95.22%	[90,42,43, 23.603016, 60.3, 6.7, 140.91]	Jute
Random Forest	98.63%	[90,42,43, 23.603016, 60.3,6.7,410.91]	Coffee
K- Nearest Neighbour	95.45%	[90,42,43, 23.603016, 60.3,6.7,410.91]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

Table 4.1: Comparison of Algorithms

Highest accuracy among the above independent machine learning algorithms were found to be 98.63the lowest accuracy was found to be approximately 10). 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 98.86output/recommendation for the given input data points of N, P, K, temperature, pH value, rainfall

Decision trees are intuitive for understanding and explanation purposes. The preparation of data is a straightforward task for users in the case of decision trees unlike most of the other classification algorithms[19]. The decision tree is made up of nodes, with the root node serving as a representation for all the rows in a dataset. Then, using a splitting variable, each node is divided into two nodes (also known as child nodes). This partitioning process is recursive. Terminal or leaf nodes are nodes that do not have child nodes. The target variable's values are carried by leaf nodes. A DT model's key benefit is that it can enable non-technical individuals to grasp the scope of a certain issue. However, the fundamental drawback of a DT model is that each node is locally optimized rather than the entire tree being optimized globally[9]. 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 [13].

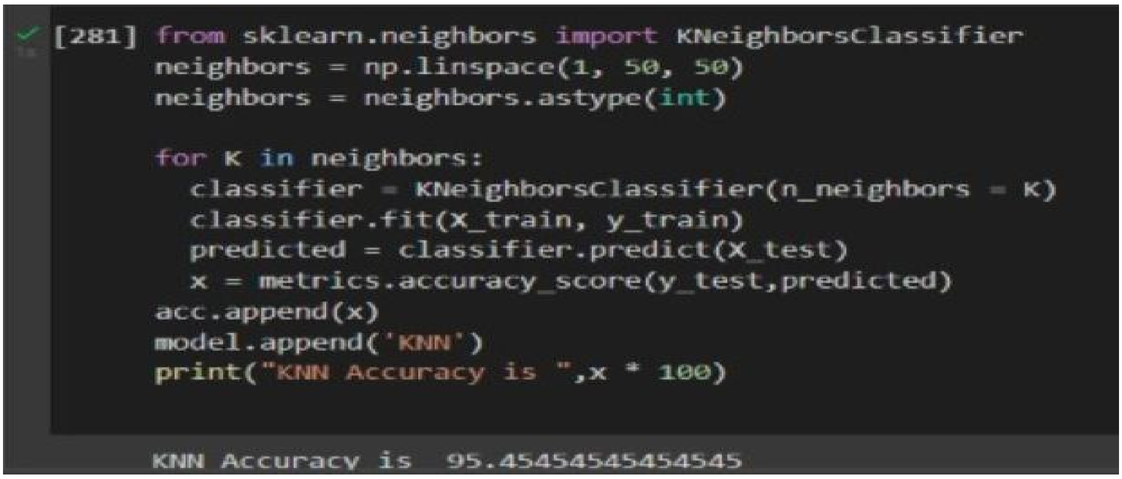


```
[17] from sklearn.tree import DecisionTreeClassifier
     DecisionTree = DecisionTreeClassifier(criterion = 'entropy',max_depth = 5,random_state = 0)
     DecisionTree.fit(X_train,y_train)
     predicted = DecisionTree.predict(X_test)
     \
     x = metrics.accuracy_score(y_test,predicted)
     acc.append(x)
     model.append('Decision Tree')
     print("Decision Tree's accuracy is", x * 100)

     # print(classification_report(y_test,predicted))
```

```
Decision Tree's accuracy is 90.45454545454545
```

Figure 4.3: Decision Tree

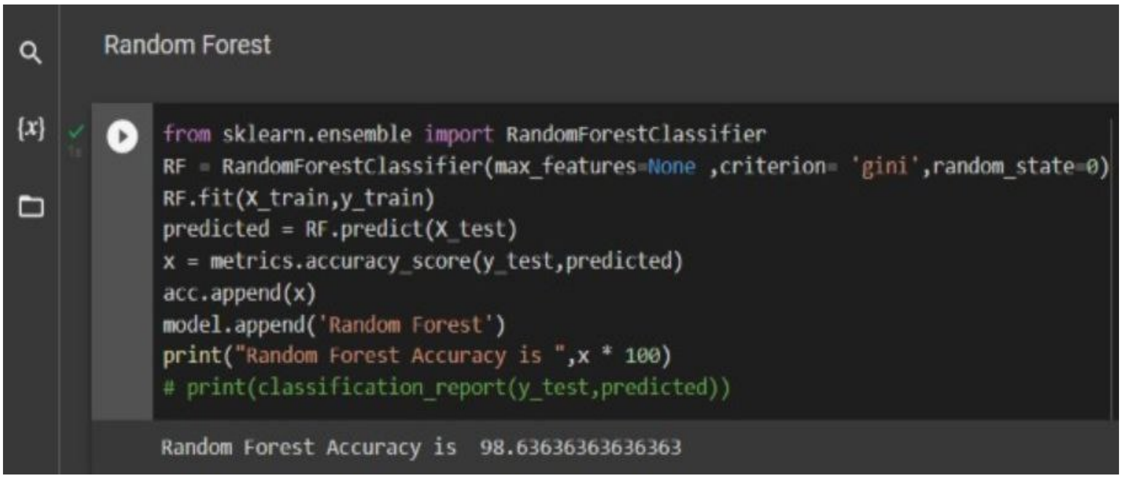
KNN was chosen because of its ability to work well with small datasets[10]. A few challenges in individual machine algorithms used above are that in KNN generally, there are four difficult problems: K computation, nearest neighbor selection, nearest neighbor search, and classification rule[6]. Given 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[15]. To deal with the lazy part of KNN classification, there are recently many efforts to set the K value or search K nearest neighbors. [17].



```
[281] from sklearn.neighbors import KNeighborsClassifier
      neighbors = np.linspace(1, 50, 50)
      neighbors = neighbors.astype(int)

      for K in neighbors:
        classifier = KNeighborsClassifier(n_neighbors = K)
        classifier.fit(X_train, y_train)
        predicted = classifier.predict(X_test)
        x = metrics.accuracy_score(y_test,predicted)
      acc.append(x)
      model.append('KNN')
      print("KNN Accuracy is ",x * 100)

      KNN Accuracy is  95.45454545454545
```

Figure 4.4: KNN

Random Forest is an ensemble method that combines the outcomes of several decision trees to make a forecast. In a Random Forest, there are two parameters: L, which determines the size of the ensemble, and tries, which indicates the number of distinct characteristics that are randomly chosen at each node. An algorithm is used to build each tree in the forest[20].Contrary to the appealing practical performance of RFS(Random Forest) in numerous real-world applications, their theoretical properties are still being studied and are not yet fully understood. 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[11].



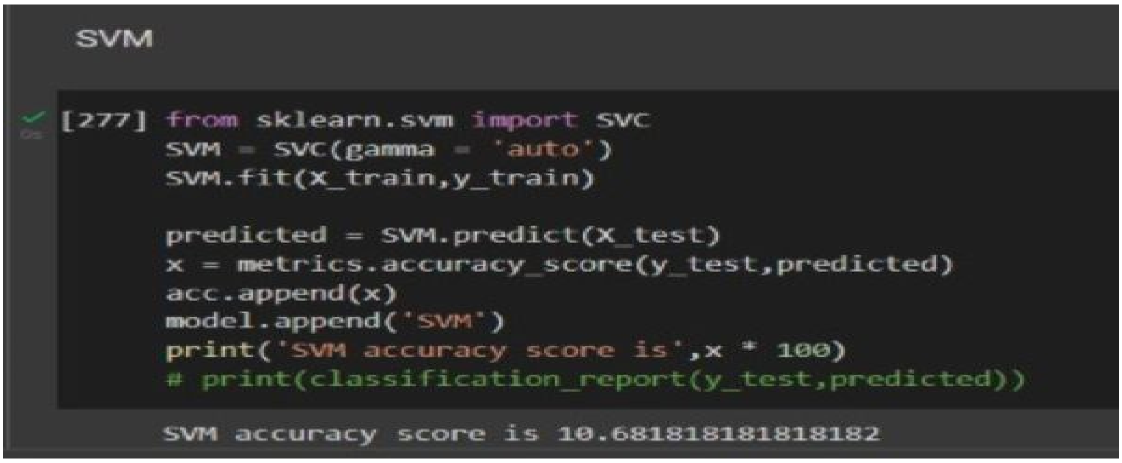
```
from sklearn.ensemble import RandomForestClassifier
RF = RandomForestClassifier(max_features=None ,criterion= 'gini',random_state=0)
RF.fit(X_train,y_train)
predicted = RF.predict(X_test)
x = metrics.accuracy_score(y_test,predicted)
acc.append(x)
model.append('Random Forest')
print("Random Forest Accuracy is ",x * 100)
# print(classification_report(y_test,predicted))
```

```
Random Forest Accuracy is  98.63636363636363
```

Figure 4.5: Random Forest

For SVM, It becomes difficult to imagine when the number of features exceeds three, and in our case, the feature is six. A technique called naive Bayes probability uses a data set from past events to estimate the likelihood of future ones[18]. In a wide range of various applications, the Naive Bayes classifier (NBC) is frequently employed in machine learning and data mining. 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[12].

The assumption of predictor independence is one of the primary causes behind the Nave 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

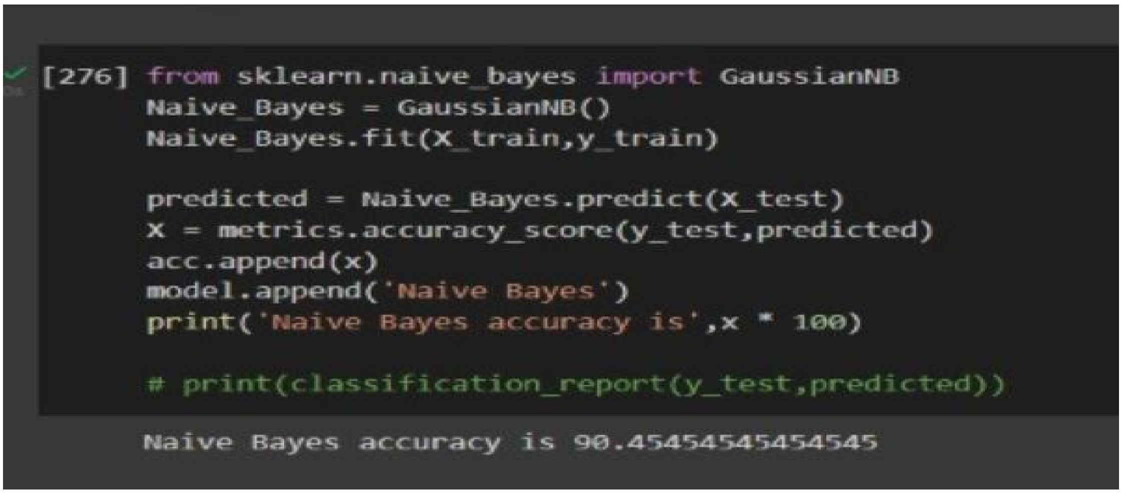


SVM

```
[277] from sklearn.svm import SVC
      SVM = SVC(gamma = 'auto')
      SVM.fit(X_train,y_train)

      predicted = SVM.predict(X_test)
      x = metrics.accuracy_score(y_test,predicted)
      acc.append(x)
      model.append('SVM')
      print('SVM accuracy score is',x * 100)
      # print(classification_report(y_test,predicted))

      SVM accuracy score is 10.681818181818182
```

Figure 4.6: SVM

under the assumption that predictors are independent[14].

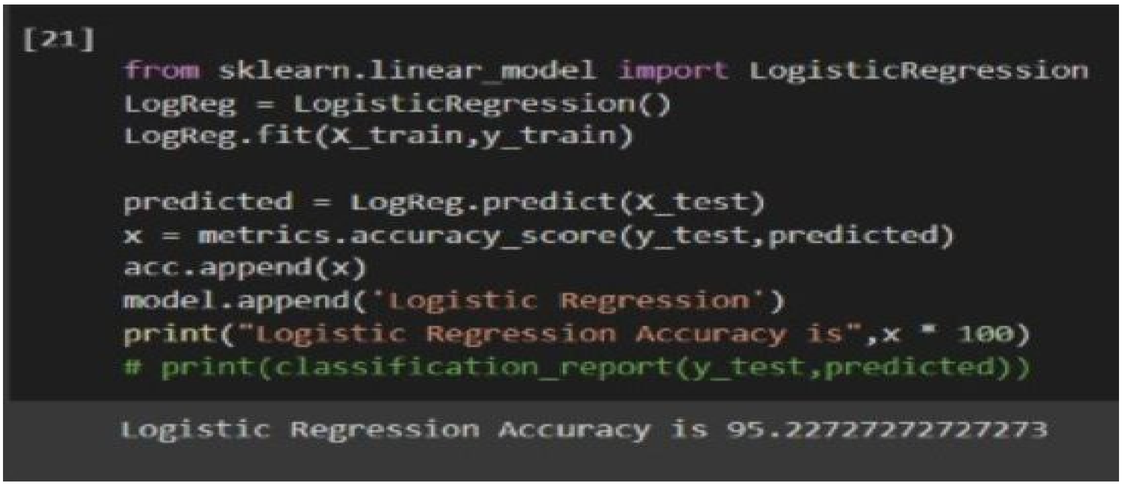


```
[276] from sklearn.naive_bayes import GaussianNB
      Naive_Bayes = GaussianNB()
      Naive_Bayes.fit(X_train,y_train)

      predicted = Naive_Bayes.predict(X_test)
      x = metrics.accuracy_score(y_test,predicted)
      acc.append(x)
      model.append('Naive Bayes')
      print('Naive Bayes accuracy is',x * 100)

      # print(classification_report(y_test,predicted))

      Naive Bayes accuracy is 90.45454545454545
```

Figure 4.7: Navie Bayes

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Each of the machine algorithms has got its own set of advantages and disadvantages. Hence, by using the majority voting classifier we aim to



```
[21]
    from sklearn.linear_model import LogisticRegression
    LogReg = LogisticRegression()
    LogReg.fit(X_train,y_train)

    predicted = LogReg.predict(X_test)
    x = metrics.accuracy_score(y_test,predicted)
    acc.append(x)
    model.append('Logistic Regression')
    print("Logistic Regression Accuracy is",x * 100)
    # print(classification_report(y_test,predicted))

    Logistic Regression Accuracy is 95.22727272727273
```

Figure 4.8: Logistic Regression

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, Naïve Bayes, Decision Trees resulting in a prediction that is closest and most suitable to give input.

The voting classifier is a type of machine learning estimator that develops a number of base models or estimators and makes predictions based on averaging their results. 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. [8].

To implement the voting classifier for our system we need to import it

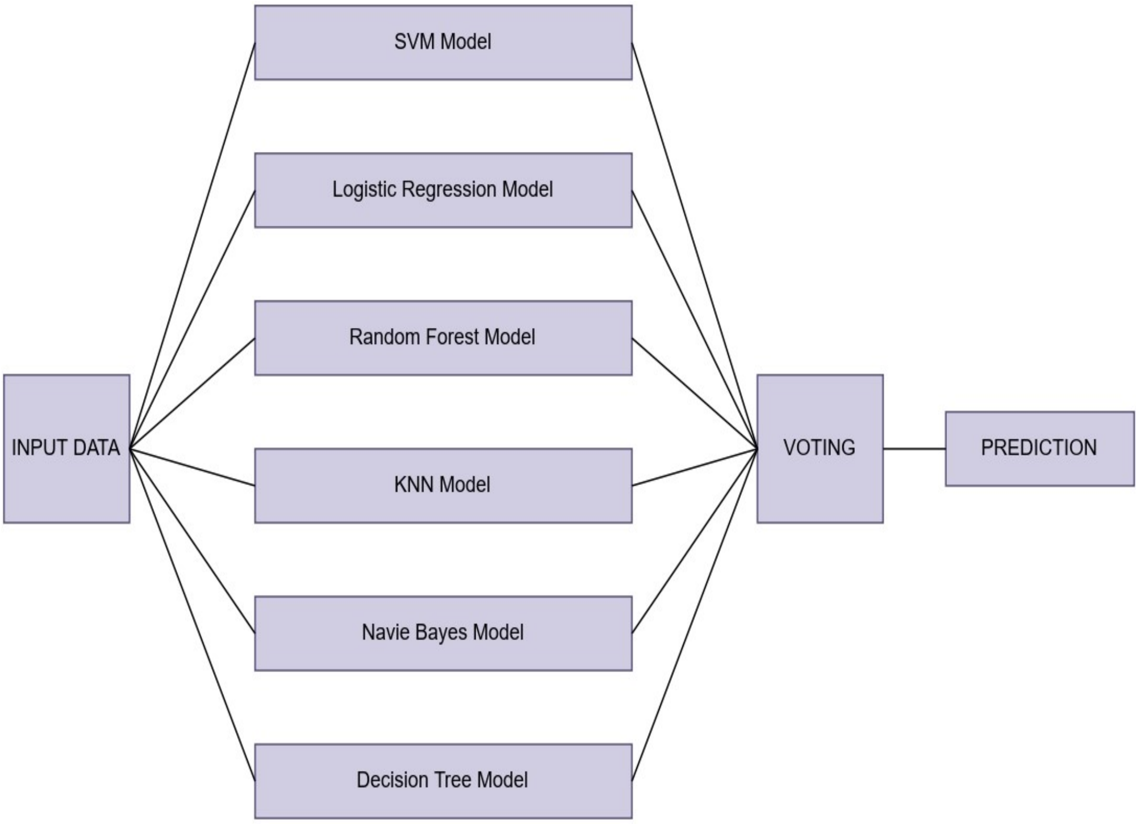


Figure 4.9: Voting Classifier

first from sckit-learn and then the below mentioned code is to be used such that all the individual models (Support Vector Machine, Naïve Bayes, Decision tree, Random Forest, KNN, and Logistic Regression) is given to the ensemble model. Based on the input data the ensemble model chooses or selects the result of the most accurate model and gives the final decision.

The concept is to build a single model that learns from these models and predicts output based on their combined majority of voting for each output class, rather than building separate dedicated models and determining the accuracy for each of them. In general, the ensemble method works fairly well when there are few positive instances and a large set of negative instances, which are also called imbalanced data sets[7]. 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.

4.5 Advantages

- Precision Agriculture could help in increasing the crop produce or yield.
- Farmers would get maximum profit by growing suitable crop due to recommendation based on certain dependent factors.
- Lessens the stress on farmers regarding the right time and conditions to grow a certain suited crop.
- Variety of crops can be grown (if there exists required conditions) and this would also help in maintain the fertility of the soil.

4.6 Disadvantages

- Any sudden Natural Disaster could impact the dependent whether and land factors which in this case, precision might not be as accurate as desired.
- Time to Time soil analysis is required to keep a check and updating the moisture and pH values of soil at certain location.

4.7 Requirement

- Dataset with soil, whether, pH values of crops and the places they are grown in India.
- Knowledge Of Python Libraries and various machine learning models to be used.

- Development Environment on systems to work with machine learning and python libraries.

4.8 Conclusion And Future Scope

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. Furthermore, using 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. And this remains our future scope

Bibliography

[1] Y. Gandge, 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), IEEE, 2017, pp. 420–423.

[2] S. Sahu, M. Chawla, 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), IEEE, 2017, pp. 53–57.