Crop Prediction Using Various Algorithms

UnnatKumar Shah
Computer Science Engineering
Nirma University
Gujarat, India
20bce301@nirmauni.ac.in

Mufid Vahora
Computer Science Engineering
Nirma University
Gujarat, India
20bce307@nirmauni.ac.in

Yati Tank
Computer Science Engineering
Nirma University
Gujarat, India
20bce323@nirmauni.ac.in

Abstract*—Machine Learning is well-equipped to analyse data on soil variables such as moisture level, temperatures, and molecular makeup, all of which affect crop growth and animal well-being. In these system we have used dataset of Potassium, Nitrogen, Phosphorus, Humidity, pH, temperature and rainfall to predict the crop yield.Using this we can develop means to predict even harvest yields and evaluate crop quality for individual species of plants to detect diseases in crop and weed infestations. Selected Machine Learning techniques such as Support Vector Classifier, Random Forest (RF), K-Nearest Neighbour (KNN), Gradient Boosting Algorithm are used in the prediction of the yield of the crop. The best results, among the contenders, were shown by the Gradient Boosting Algorithm with 99.5% accuracy.

***Keywords*—Support Vector Classifier, Random Forest (RF), K-Nearest Neighbour (KNN), Machine Learning, Crop Prediction, Analysis of features, Gradient Boosting**

I. Introduction

Today, technology can allow harvests to be cultivated with far more precision, allowing growers to treat organisms practically individually, considerably enhancing the efficiency of farmers' decision-making. This research suggests a practical and user-friendly yield forecast method for farmers.

1.1 Motivation

In this work, we aim to expand the accuracy of the result of the crop yield prediction algorithm. In this paper, we are going to implement different machine learning algorithms for Crop Yield Prediction which are KNN, SVC, Random Forest, Decision Tree and Gradient Boosting. Then, we will be comparing them to see which one is better to classify data. As we all know, Crop Yield Prediction is so crucial nowadays as it can help prevent drought and disaster in the future .

1.2 Contribution

Machine learning offers numerous opportunities for research and development in a variety of fields, including crop yield prediction system, recommendation system, and by using algorithms, we get the best outcomes. By examining the terms they include, we can determine which crop can be grown on the certain land.

1.3 Problem statements

For global food production, crop yield forecasts are essential. Import and export choices may be made quickly using an efficient strategy to increase nutritional security. To introduce better varieties for a range of environments, seed companies must forecast new hybrids. Crop forecasts help farmers make money by helping them avoid financial losses, plan their planting seasons, and take the appropriate action depending on the soil and environmental dynamics.

The urgent requirement is to make a system that might provide Indian farmers with predictive information so they could carefully select which crop to farm with knowledge.

II. METHODOLOGY

In the term paper, classification such as KNN, Support Vector Classifier(SVC), Decision Tree, Random Forest and Gradient Boosting is used to classify data and is then analyzed.

A. KNN

K-NN (k-Nearest Neighbors) algorithm comes under supervised learning. This can be used for classification as well as regression approaches.

As K-NN engages the odd k number, it is referred to as the lazy learner technique in machine learning. K nearest neighbors is a simple method that helps in the classification of recent information based on a correlation and retains all provided data.

B. SVM

The SVM algorithm's goal is to put in place the best line or distance measure that can split n-dimensional space into various classes, allowing us to quickly differentiate new data points in future. This best decision boundary is called a hyperplane.

SVM selects the ultimate vectors and points that aid in the hyperplane creation. Support vectors, which are used to show these extreme instances, form the roots for the SVM method.

C. Decision Tree

The decision tree is a non-parametric supervised learning technique that is applied to classification and regression issues. It gets a base node, branches, internal, and leaf nodes and is arranged hierarchically. There are 3 type of Decision Tree Algorithms - ID3, C4.5 and CART.

D. Random Forest

As per the name suggests, a random forest is made of a variety of single decision trees that toil jointly as a group. The class with the most votes becomes the suggestion given by the model. The random forest's individual trees each split forth a prediction of the class.

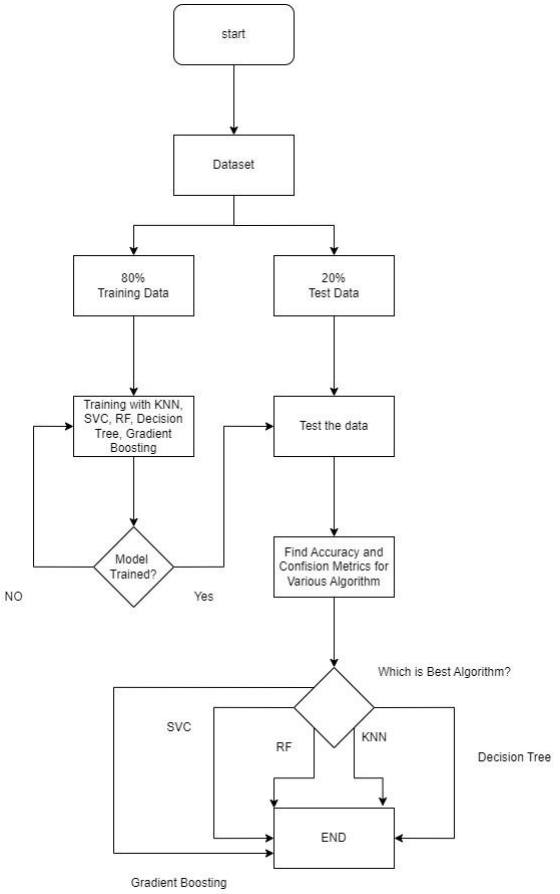
E. Gradient Boosting

A machine learning technique known as gradient boosting put in effect, among different stuff, for classification and regression duties. It gives a prediction model in terms of a variety of weak prediction models such as the decision tree. The resultant technique, known as gradient-boosted tree, usually outperforms random forest whenever a decision tree is the poor learner.

III. Solution Statement

For this project the data set used here is **Crop_recommendation.csv[1].** The data set contains 2201 rows and 7 columns where different features such as N,K,P, temperature,humidity,pH, and rainfall of different crops were used.

In the next step we have imported the pandas library and by using pandas we have read the csv file and copied it into a dataframe. In the next step we have used Heatmap to check all the null values present in the dataset.



IV. Algorithms

1)KNN

Steps for classification using knn algorithm

1)Collect the data then identify the missing data,encode the categorical data ,apply feature scaling and then split the dataset into 80:20 ratio for training set and testing set respectively

2)Choose the optimal value of k,generally optimal value of k is square-root of number of features in the dataset

3)Calculate the distance of neighbors by using Euclidean distance.

4)Determine K nearest neighbors.

5)Classify the testing data according to their similarity with the k nearest neighbor.

2)SVM

Support Vector Machine creates a hyperplane that can separates the features from each other so that if there is new data point it can be identified in its respective class , extreme points in of the features are used to create a hyperplane they are known as support vectors,margins are the distance between support vector and hyperplane. Hence this algorithm is known as a support vector machine ,this classification algorithm depends upon the hyperplane which differentiates the classes.

3)Decision Tree

Decision tree algorithm is a graphical tree structured classifier where the tree is divided upon the features and branches of the tree are decisions which have to be made to achieve the solution . Steps for decision tree are,

1)Collect the data then identify the missing data,encode the categorical data ,apply feature scaling and then split the dataset into 80:20 ratio for training set and testing set respectively

2)Root node has to be selected which justifies our dataset or is the best representation of our dataset. The method used for selecting the dataset is gini index , the attribute with the lowest gini index should be chosen as the root node.

3) Further the tree should be divided upon the lowest gini index and repeat this procedure until all the attributes are used as the nodes of the tree and no leaf node can be classified

4) Classification of new data points is made as per the decision tree created.

4)Random Forest

Forest stands for the collection of the decision trees and random because some number of points are chosen randomly for the samples from the training set. By using decision tree problems of overfitting occur hence we use random forest algorithm, here we create multiple decision trees .Steps for random forest are

1)Collect the data then identify the missing data,encode the categorical data ,apply feature scaling and then split the dataset into 80:20 ratio for training set and testing set respectively

2)Random samples are selected from the training data.

3)Decision tree is created from the random samples selected as discussed earlier gini index is used to create the decision tree.

4)Voting is done by averaging the decision trees created from random samples.

5)Classification of decision trees is done by majority voting which means we will find the prediction for new data points and then will classify it as per majority of votes obtained.

4)Gradient Boosting

The main idea behind this approach is to build models one at a time, each one aiming to improve on the shortcomings of the model before it.

The aim is to minimize this loss function by adding weak learners using gradient descent. We will have several loss functions addressing regression concerns because it is dependent on a loss function, we will have many loss functions for categorizing, like Mean Squared Error (MSE).

V. PERFORMANCE METRICS

A. Parameters

1. KNN

K-values	1-10
metric	minkowski

2. SVM

C	0.1
Kernel	rbf, linear, poly
Data type Features	int
degree	3
gamma	scale

3. Decision Tree

Random State	42
Criterion	Gini
Data type Features	int

4. Random Forest

Random State	42
Criterion	Gini
max_depth	4
n_estimators	100
Data type Features	int

5. Gradient Boosting

Loss	log_loss
Criterion	friedman_mse
Learning rate	0.1
n_estimators	100

B. Test Beds

For the Crop Yield Prediction problem, the algorithm was implemented in google colab using python programming

language and various libraries of python such as pandas for reading csv file, Numpy for numerical calculations,sklearn for implementing various algorithms. Also for plotting graphs we have used the SNS library.

C. Performance Metrics

Algorithms	**Accuracy**
KNN	0.9781818181818182
SVC- Linear Kernel Accuracy Rbf Kernel Accuracy Poly Kernel Accuracy After Parameter Tuning	 0.9745454545454545 0.9872727272727273 0.9890909090909091 0.9866710547967747
Decision Tree	0.9872727272727273
Random Forest	0.97
Gradient Boosting	**0.9927272727272727**

D. Result Analysis

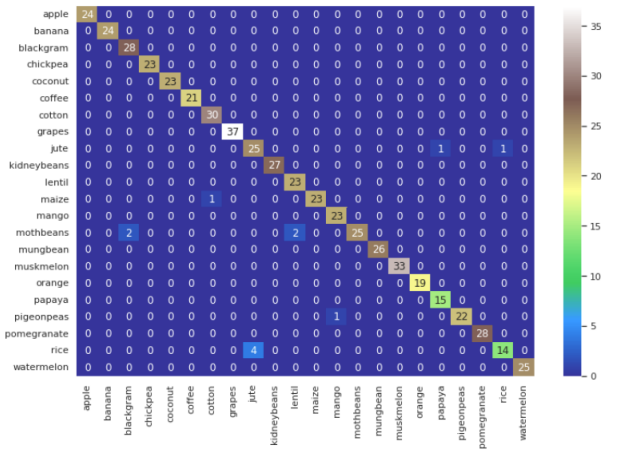
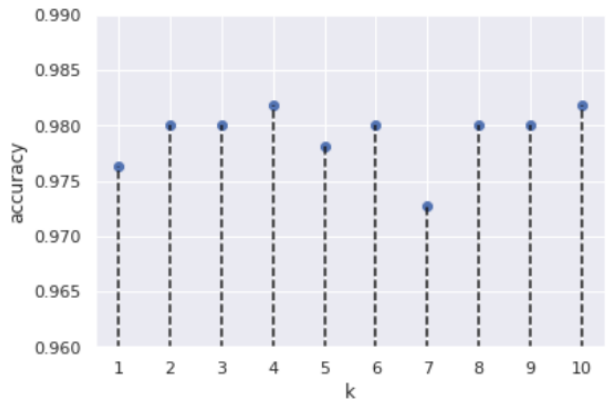


Fig 5.1

This graph shows us the plot of the confusion matrix used to evaluate the performance of the classification algorithm. It is the count of predicted value is actual value , when it is actually apple 24 times it has predicted apple ,for blackgram it has 28 times predicted blackgrams and 2 times predicted moth beans , this is analysis for KNN model



Fig, 5.2

Accurate K-value can produce a best-fit model , as k-value is changed the accuracy of the model changes , hence choosing optimal value of k is to be choosed. Here the graph shows accuray as the value of k changes. Highest accuracy is achieved when the value of k is 4

It is observed that the value of k should be the square root of the total number of features as we have 22 features , 4 would be the optimal value of k.

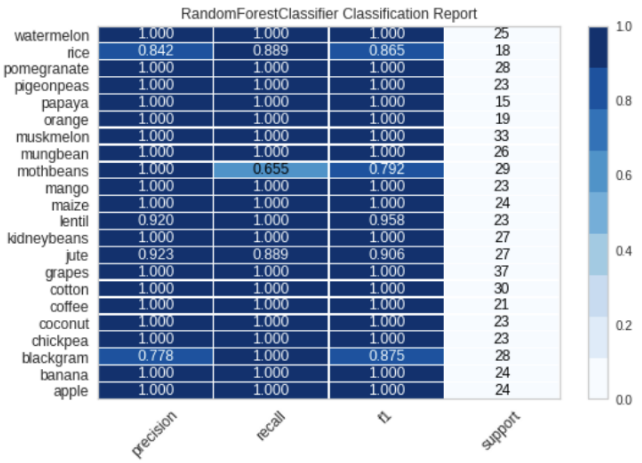


Fig 5.3

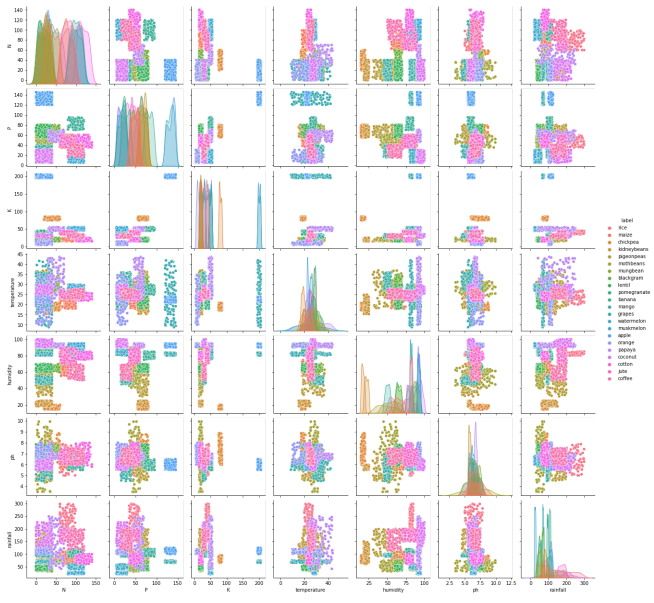


Fig 5.4

This plot is used to represent the relationship between the variables in the dataset, the diagonal plot represents univariate distribution in which Nitrogen has been plotted against Nitrogen.

For non-diagonal graph it represents the effect of two variables on the crops for example rain affects the soil moisture which indirectly affects the ph level of the soil, rice and coconuts needs high rainfall and high humidity , so here is a plot for rainfall and humidity which shows that rice and coconut are more likely to be produced in the circumstances of high rainfall and high humidity

.

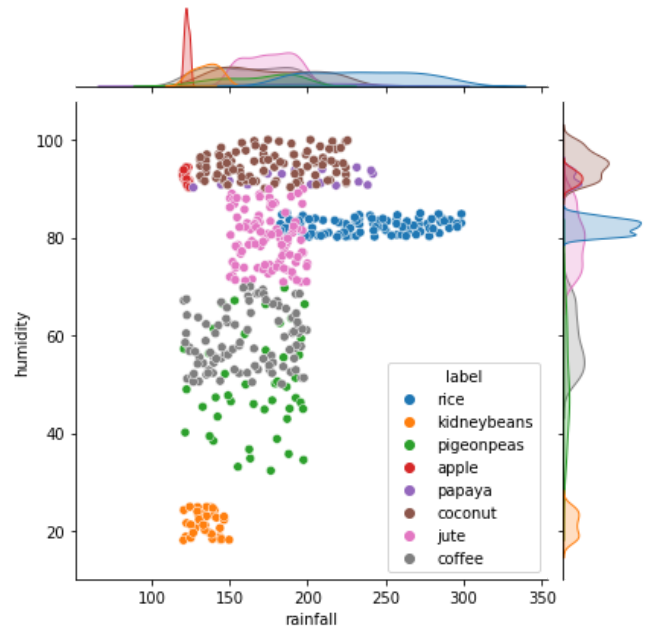


Figure 5.5 Plot of Rainfall vs Humidity

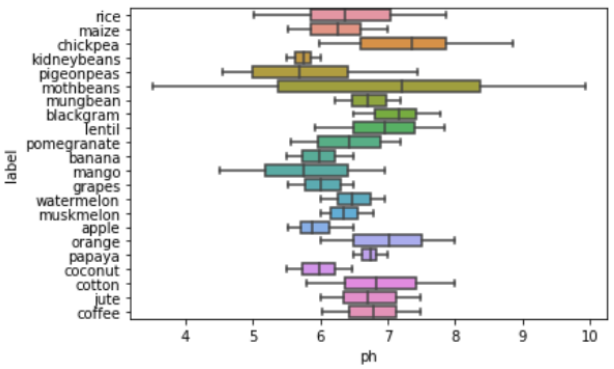


Fig 5.6

This graph shows the ph value required for the crop as we can observe that the best ph value is between 6 and 7. All the crops lie between this ph value of soil , this shows us that neutral value of ph should be achieved to grow crops.

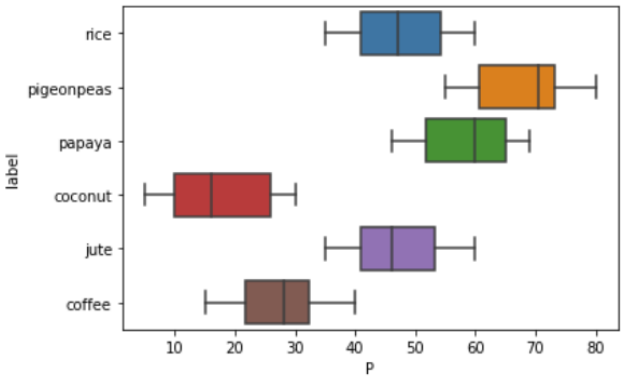


Fig 5.7

This graph shows us the phosphorus levels of different crops when rainfall fall is above 150 mm as we can observe that phosphorus levels differ at a significant scale in give prerequisite.

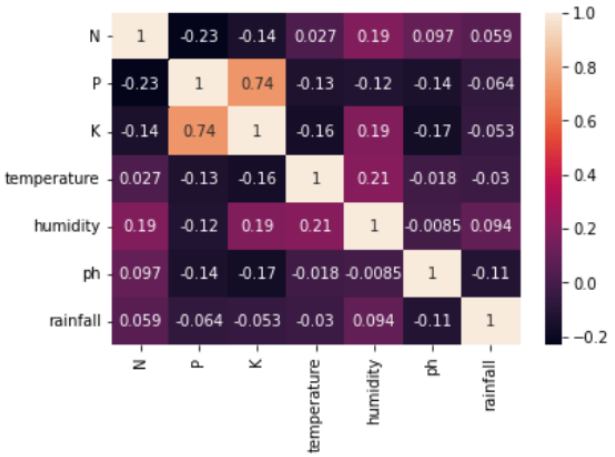


Fig 5.8

This is heatmap representation of all the features on x and y axis , here the diagonal is used to represent that feature is related to itself hence all the diagonal values are 1, in non-diagonal value the it represents the one feature affects the other as we can interpret from the graph that Potassium and Phosphorus are highly dependent on each other.

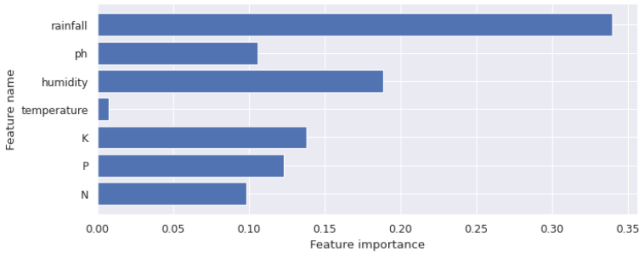


Fig 5.9

This graph represents the feature name and its importance.As we have used the gini index for the classification using the decision tree , an important feature has been identified and splitting of the tree is based upon that important feature, so here the most important feature is rainfall which further is divided as per the humidity level and so on.

V. CONCLUSION

In this article, We have used machine learning algorithms for predicting the crop yield system which are SVC, KNN, Decision Tree, Random Forest and Gradient Boosting. By contrasting these algorithm's accuracy, we see Gradient Boosting gives more accurate results than any other algorithms.. For Crop Yield Prediction, Gradient Boosting is a good algorithm.

VI. REFERENCES

[1] A. Dhande and R. Malik, "Empirical Study of Crop-disease Detection and Crop-yield Analysis Systems: A Statistical View," 2022 International Conference on Emerging Smart Computing and Informatics (ESCI), 2022, pp. 1-4, doi: 10.1109/ESCI53509.2022.9758284.

[2] M. Chandraprabha and R. K. Dhanaraj, "Soil Based Prediction for Crop Yield using Predictive Analytics," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 265-270, doi: 10.1109/ICAC3N53548.2021.9725758.

[3] P. Saini and B. Nagpal, "Efficient Crop Yield Prediction of Kharif Crop using Deep Neural Network," 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 2022, pp. 376-380, doi: 10.1109/CISES54857.2022.9844369.

[4] S. K. Sharma, D. P. Sharma and J. K. Verma, "Study on Machine-Learning Algorithms in Crop Yield Predictions specific to Indian Agricultural Contexts," 2021 International Conference on Computational Performance Evaluation (ComPE), 2021, pp. 155-166, doi: 10.1109/ComPE53109.2021.9752260.

[5] Z. Doshi, S. Nadkarni, R. Agrawal and N. Shah, "AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697349.