Finding Recommendation from Crop Dataset using Map Reduce Machine Learning Algorithm

Dr. Chandrashekar D K
Associate Professor
Department of Computer Science &Engineering
SJB Institute of Technology

[1] Supriya Kumari
Department of Computer Science & Engineering
SJB Institute of Technology

[2] Yash Raj Mishra
Department of Computer Science & Engineering
SJB Institute of Technology

[3] Sagar Devanur
Department of Computer Science & Engineering
SJB Institute of Technology

[4]Vivek Kumar
Department of Computer Science & Engineering
SJB Institute of Technology

Abstract* — The country's economic prosperity and expansion are mostly attributed to agriculture. Approximately 17% of India's GDP comes from agriculture and related industries, and 70% of Indians still choose this as their primary line of work. The farmers' failure to select the proper crop for cultivation is the primary and most significant hindrance to crop productivity. Farmers can benefit greatly from this kind of technology because it is portable, simple to use, and doesn't require a lot of hardware. It becomes much more crucial in India, since the agricultural research and development agency lacks sufficient facilities. Random Forest and Naive Bayes are the independent base learners employed in the ensemble model. The crop recommendation algorithm classifies the input by separating the suggested crop types, Kharif and Rabi, from the soil dataset.. The dataset includes samples of the average rainfall and surface temperature as well as the physical and chemical properties particular to the soil. Combining the independent base learners yields an average classification accuracy of 99.91 percent.

Keywords – Map Reduce, Naïve Bayes, Random Forest, Decision Tree , crop recommendation

I. INTRODUCTION

It is anticipated that the demand for food would increase by more than 70% in the near future due to the growing population of the world. Currently, the availability of the crop is not sufficient to feed such a large population .The reason could be less production of crop,wastage due to various climatic conditions .In near time, the world would face a situation of food crisis. To overcome such unescapable situation ,the production of crop has to be increased. This needs a precise and managed farming which include the 1st step of farming that is correct crop selection for the land.

It is often observed that technology had and will always play a major role in making the life of people easy and struggle free. Here comes the smart technology of farming into picture .The huge farming system which is considered as the major pillar of the country produces complex and big data over years .These data can play a major role in the enhancement of crop production if used in smart way .Big data focuses on volume, variety as well as velocity. Then the application machine learning algorithm on these data make it really useful .The parameters considered in this paper contains different parameters related to environmental condition and soil condition.

The very first step in farming is to decide the crop for the land. Numerous variables, including the local climate, the amount of rainfall, the soil's quality, the temperature, and many others, have an impact on this process. .It is a hectic task to get the most probable crop for particular region and particular time. Considering these factors and then finding the precise crop of most accuracy will be termed as recommendation.

Machine learning is one such technology where different algorithms like Decision tree, Naive Bayes ,Random Forest and Logistic regression are of much use when dealing with the training of model .These algorithms take into account different factors and processes for training the model .Each of them has its own method of training the model using huge datasets.And when the accuracy of each prediction is compared to the others, this causes variation. This paper focuses on the accuracy measurement of each model trained under different algorithm of machine learning and find the most suitable one.

As mentioned earlier, dealing with huge dataset is a time taking process. To overcome such situation the deployment of the project is done on spark, an engine for processing large-scale datasets. The technique behind this process is called map-reduce technique. Basically, map-reduce is a technique for distributed computing. The main aim of this technique is decomposing the data in key-value pair and run in a parallel processing computation. This leads to better, enhanced and fast processing of data. The principle aim of this paper is to get the best accuracy prediction model for various input data.

II. LITERATURE SURVEY

Zeel Doshi et al. [1] , have proposed a model that takes into consideration all the factors affecting the growth of crops.The factors considered are rainfall,temperature,location,soil condition.The two subsystems are used in the process, one is crop suitability predictor and the other is rainfall predictor. Crop suitability is predicted after perfoming various machine learning algorithms on datasets.For the rainfall prediction linear regression algorithm is perfomed on datasets. Using this two models an intelligent crop recommendation system has been designed.

M.V.R Vivek et al.[2], proposed methods that deal with choosing the right crop and climate for the field. A technique known as the majority voting mechanism is utilised for crop prediction. Using machine learning methods like naive bayes and SVM, a comparison of soil categorization is conducted. It offers some pertinent research on how the data mining system for predicting the climate produces superior results and can be considered as a substitute for conventional metrological ways of predicting the climate.

A. Suruliandi et al.[3], proposed a sequential approach of predicting the crop using feature selection techniques and classifier. Important attributes have been extracted using wrapper feature selection technique. Analyzing the different performance metrics, the best selection feature and classification method is revealed.

Vaishnavi.S et al.[4], proposed a simple approach of predicting crop based on the previous production on the field in a restricted particular area. The data of the farm production is collected from previous crop yields and the model is trained using different methods like neural network and soft computing. Using these methods the best prediction is done for collected from previous crop yields and the model is trained using different methods like neural network and soft computing. Using these methods the best prediction is done for high volume of production.

G. Mariammal et al.[5], An important aspect of agriculture, crop cultivation and crop forecast mostly depend on soil, weather conditions including temperature and rainfall, and the amount of fertiliser applied, especially nitrogen and phosphorus. This machine learning feature selection aspect plays a significant role in choosing the most important features for a given location and continues the crop forecasting process. updated.

They have come to the conclusion that agriculture depends on the ability to predict a suitable crop for cultivation. A fantastic method for choosing features utilising a permutation crop data set and a rating has been proposed in this paper: multiresolution feature extraction. To determine whether machine learning algorithms, such as kNN, NB, DT, SVM, RF, and bagging classification approaches, are most effective at predicting the most suitable crops for cultivation, numerous experiments have been carried out.

Shiyam Talukder et al. [6], the most productive agriculture output requires the right crop for a certain place. Here, they have created a model that uses prediction and suggestion with various machine learning techniques to calculate production depending on the factors temperature, humidity, and precipitation. They did a contrast of the algorithms by examining the precision after training the dataset and using these techniques. On the other hand, they used Multi-Condition Filtering and Collaborative Filtering algorithms for the recommendation. The subject of this paper's research, which is based on machine learning methods for prediction, is agricultural production. As you can see, every algorithm shows that the rice harvest is the most abundant. The three most effective crops for recommendations made using collaborative filtering are rice, maize, and moong. Multi-condition logic, on the other hand, filters the crops based on the state of the crop yi productivity factors. This is how they have worked on various ideas condition logic filters the crops according to the condition of the crop yielding factors. In this way, they have worked on various ideas to develop the required model.

Mummaleti Keerthana et al.[7], Understanding practical and real-world use cases for agricultural production prediction requires an understanding of machine learning. This study examined the use and use of ensemble techniques for crop type prediction based on factors like location. Through the use of our search method, we were able to finalise 28242 occurrences and nearly 7 features across many databases. We learned that Neural Networks and Decision Trees are the more popular methods for these models from numerous foundation articles. Decision trees use parameters like maximum depth and nestimators, thus we may improve the outcomes by changing those factors. After conducting research, we came to the conclusion that an ensemble of the decision tree and AdaBoost regressors produced extremely accurate results. Crop yield forecasting includes forecasting the output of the model.

They claimed to have constructed a system for agricultural production prediction using data that had previously been gathered in the end of this research. Some machine learning techniques have been used to do this. To forecast the outcome with a higher rate of accuracy, an ensemble of decision tree regression and an AdaBoost regression is utilised in this case. Farmers will be helped to make an informed selection in selecting the best crop for production by the worldwide accuracy of crop prediction..

III. PROBLEM STATEMENT

To put forth a novel strategy using a map-reduce-based machine learning algorithm to help farmers choose crops by taking into account all the variables, such as the sowing season, soil, and geographic location to provide an accurate result using crop dataset.This new approach must be done using spark clusters for processing the big dataset so that farmers will get the accurate prediction of the crop for suitable factors as mentioned above.

IV. METHODOLOGY

A. Dataset description

This study used a set of agricultural data that was mostly composed of soil and environmental variables. The data set for environmental factors is accessible to everyone on the website www.tnau.ac.in. But the public cannot access the soil characteristics data set. As a result, it is manually gathered from a variety of places, including the Department of Agriculture in Tenkasi, India's Sankarankovil Taluk. It was created specifically for this study project. The data collection includes 2201 occurrences with 8 attributes, while the remaining 4 features are environmental parameters and the remaining 4 features are soil characteristics.

The types and descriptions of the soil and environmental factors that affect crop forecast are shown in Table I. About 22 different types of crops, including rice, maize, chickpeas, kidney beans, pigeonpeas, moth beans, mung beans, blackgram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee, are included in the target value.

Serial No.	Attribute	Description	Data type
1	Temperature	Temperature of a place.	Double
2	Humidity	Humidity of a place	Double
3	Rainfall	Altitude of rainfall	Double
4	PH	Acidity or basicity of the soil	Double
5	label	Target crop name	String

Table 1 Data Collection description

The environmental factors data set is accessible to everyone on the website www.tnau.ac.in.On the other hand, the public cannot access the soil characteristics data set. As a result, it is manually gathered from a variety of places, including the Department of Agriculture in Tenkasi, India's Sankarankovil Taluk. It was created specifically for this study project.

1000 cases with 9 classes and 16 features total in the data set, 12 of which are soil characteristics with the remaining 4 being environmental factors.

B. Classification Technique:

The fundamental learning process in machine learning (ML) is classification, which predicts the target class of an input. supervised and unsupervised classification techniques are the two categories. NB, DT, RF, and Logistic Regression are examples of supervised learning techniques used in this study.

1. **Naive Bayes**

A straightforward classification system called the NB classifier [4] selects the crop with the highest probability by estimating the probability of each class. To determine the best crop for cultivation, the NB classifier is trained using training samples, and its performance is assessed using testing samples from the testing set.

2. **Decision Tree**

A supervised learning model with a structure like a tree, the DT. Each internal node is top-down and is labelled with an input feature [5]. The class that was utilised to forecast the target variable is labelled on each leaf node. Tree splitting is significant for the DT, which contains the prediction class. The testing set's data values are isolated and used separately. to determine the best crop.

3. **Random Forest**

A group of binary DTs make up the RF classifier [7]. The RF creates a collection of DTs from a portion of the training data that is selected at random.. For crop prediction, each individual tree divides a class according to the Gini index valueA new sample from the testing set is categorised based on the votes cast by the majority of the trees in the RF. [15].

4.**Logistic Regression**

A logistic regression equation can be used to estimate probabilities in order to analyse the relationship between the dependent variable and one or more independent variables. statistical software. In regression analysis, logistic regression is used to estimate a logistic model's parameters. Most frequently, maximum-likelihood estimation is used to determine a logistic regression's parameters.

5.**Map Reduce technique:**

Java-based MapReduce is a processing method and a model for a distributed computing programme. Two crucial tasks, Map and Reduce, are part of the MapReduce algorithm. Map transforms a set of data into another set of data in which each component is divided into tuples. The second work is a reduce task, which takes a map's output as input and concatenates the data tuples into a smaller collection of tuples. The reduction work is always carried out following the map job, as the name MapReduce implies.

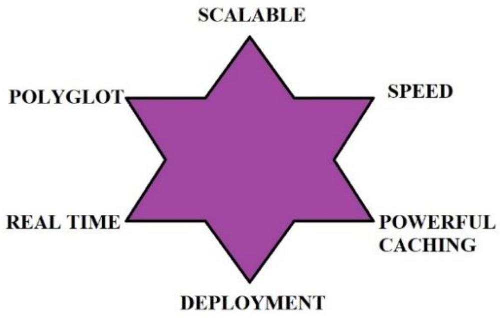


Fig. 1 Spark Properties

Scalable

A system's capacity to manage an increasing volume of work by adding resources is known as scalability. A scalable business model means that a corporation can increase revenues given more resources in an economic situation. A package delivery system, for instance, can deliver more packages by adding additional delivery cars.

Polyglot

High-level APIs for Java, Scala, Python, and R are offered by Spark. These four languages are all capable of producing Spark code. Additionally, it offers Python and Scala shells. Scala, R, Python, Java, and Clojure are just a few of the languages that can be used to create Spark applications.

Powerful caching

A straightforward programming layer offers effective cache and disc persistence features.

Deployment

It can be set up using Mesos, YARN for Hadoop, or Spark's built-in cluster management.

Real-Time

It offers real-time processing and reduced latency because the calculations are done in memory.

Speed

Spark is up to 100 times faster at processing large volumes of data than Hadoop MapReduce. This performance can also be achieved through careful partitioning.

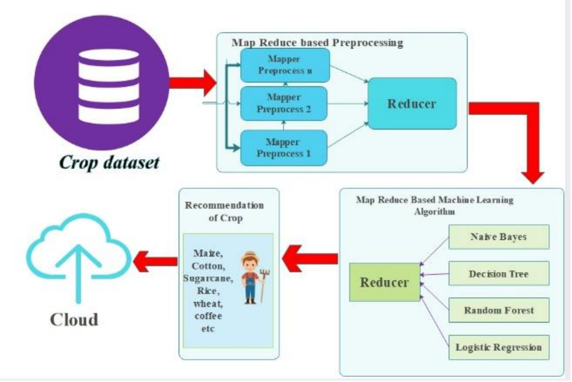


Fig.2 Map Reduce based system Architecture

Map Reduce Preprocessing Steps:

- The map accepts pairs of data and outputs a list of "key, value" pairs. In this situation, the keys won't be distinctive.
- The Hadoop architecture employs sorting and shuffling based on Map's output. This sort and shuffle operates on this list of "key, value" pairs and sends out "key, list(values)" pairs of unique keys and lists of values corresponding to those keys.
- The reducer phase received a sort and shuffle output. On a list of values for distinct keys, the reducer applies a predefined function, and the final result, "key, value," is stored or shown.

Prior to the reduction, sorting and shuffling take place on the output of the mapper. The results are sorted by key after the Mapper process is finished, divided if there are several reducers, and then written to disc. We gather all the values for each distinct key k2 using the input from each Mapper k2,v2>. The reducer phase receives this output from the shuffle phase as an input in the form of k2, list(v2).

Hadoop's two main parts, MapReduce and HDFS, are what make it so strong and effective to utilise. A programming model called MapReduce is used for distributed, parallel processing of massive data collections. To create the final output, the data is separated first, then combined. There are numerous programming languages used to create the libraries for MapReduce, each with its own unique set of optimizations.

Cloud

Because computation and storage occur on servers in a data centre rather than locally on the user device, users can access the same files and programmes through the cloud from nearly any device. Similar to this, after preprocessing, the entire dataset is kept in the cloud for future use and project improvement. Data is being prepared for use by the cloud for upcoming performance calculations.

Components of MapReduce Architecture:

Client: The Job is brought to the MapReduce for processing by the MapReduce client. There may be a number of clients accessible that continuously send jobs to the Hadoop MapReduce Manager for processing.

Position: MapReduce Job refers to the actual job that the client requested to be completed and is made up of numerous smaller tasks that the client desires to process or carry out.

Hadoop MapReduce Master: It separates the specific task into several task-parts.

Job Parts: The duties or sub-jobs that result from splitting the primary job. the end product produced when all the job-parts are integrated.

Input data: The data set provided to MapReduce for processing is known as the input data.

Output Data: After processing, the ultimate outcome is discovered.

C. Developing the Prediction Model:

Steps for crop prediction are mentioned below:

Step 1: The crop data set that includes soil and environmental characteristics is the input data set.
Step 2: The input dataset undergoes preprocessing to normalise the data. Anomalies are checked for in redundant data and missing values.
To maintain the data set consistent, variables in the dataset are also translated into a specific range.
Step 3:To determine the prediction's accuracy, proposed techniques and machine learning algorithms like Map-Reduce are employed. And a solo local machine is used to run the project.
Step 4: To launch the project, create a spark cluster in Google Cloud or Amazon E2C.
Step 5: The project is deployed on the Spark cluster quickly while taking the time factor into account.

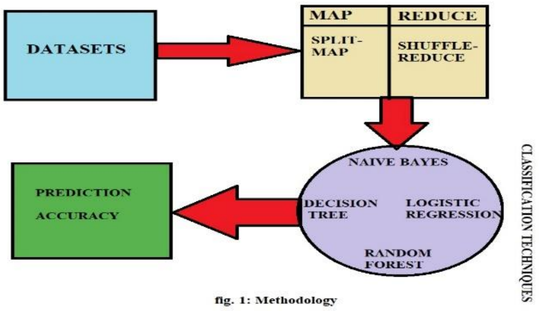


Fig.3 Methodology

Map

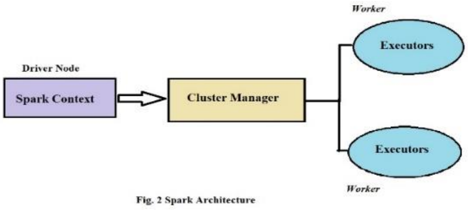
Every element of an RDD or DataFrame can have the transformation function (lambda) applied to it using PySpark map (map()), which creates a new RDD as a result. You will discover how to use the RDD map() transformation with a DataFrame as well as its syntax and usage in this tutorial.
Any sophisticated actions, such as adding a column, updating a column, modifying the data, etc., are applied using the RDD map() transformation; the output of a map transformation will always have the same amount of records as the input.

Reduce

Reduce is a spark action that uses a function to aggregate a data set (RDD) element. That function returns one after accepting two parameters. (Function | Operator | Map | Mapping | Transformation | Method | Rule | Task | Subroutine) must be enabled for the function.

Datasets

There is a need to choose important features that aid in identifying appropriate crops for particular sections of land because the produced dataset contains a variety of traits and characteristics in an unordered way and the parameters vary for every zone crop field.



Fig,4 Spark Architecture

Apache Spark

An open-source unified analytics engine for analysing enormous amounts of data is Apache Spark. An interface called Spark allows clusters to be programmed with implicit data parallelism and fault tolerance. When used alone or in conjunction with other distributed computing tools, Apache Spark is a data processing framework that can quickly conduct operations on very large data sets and distribute operations across several machines.

Cluster manager Spark can be run on the Cluster Manager platform (in cluster mode). Simply said, the cluster manager manages all nodes in accordance with needs and supplies resources to all worker nodes as necessary. In a cluster, we can state that a master node and worker nodes are available.

The table above displays the observations that our machine-learning-based algorithms in databricks, Apache Spark's web-based platform, were able to account for. These observed results allow us to determine which machine learning method has the highest accuracy and shortest time period, allowing us to apply that specific algorithm for crop prediction.

V. RESULTS

Algorithm	Accuracy	Time (in secs) (Local Machine)	Time (in secs) (Apache Spark)
Decision Tree	0.44593	0.30284070 96862793	0.196209192 27600098
Random Forest	0.92678	1.40250778 19824219	0.562893727 1890073
Naïve Bayes	0.88354	0.55456852 91290283	0.079200201 91828390
Logistic Regression	0.66345	0.98506116 86706543	0.099210838 38444400

Table 2 Accuracy and Time Comparison of

Different Algorithms

In this article, we have examined the speed and accuracy of the various algorithms decision tree, naive bayes, logistic regression, and random forest. The programme analyses several time- and accuracy-related algorithms. When the code is performed on a local system without Spark and subsequently in an Apache Spark cluster, the accuracy for various algorithms is reported. The inference is made that the cluster runs the algorithm and efficiently determines the outcome in a very short amount of time.

Measures of performance By adjusting the total number of trees' splitting range and analysing the ranking mechanism used to determine each feature's importance, the suggested technique's performance is assessed. In this work, the metrics of ACC, precision (P), recall (R), specificity (S), and F1 score are used to assess the effectiveness of the FS and classification algorithms.

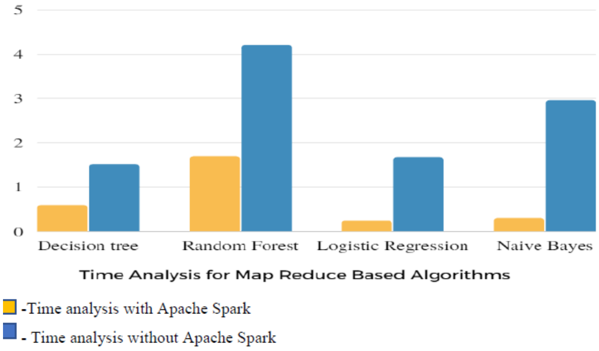
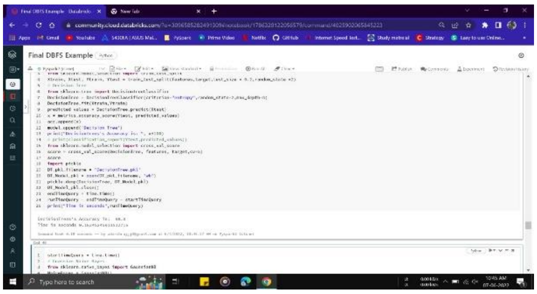


Fig. 5.1 Time analysis for Map Reduce based algorithms

Four stages were included in the data collection, algorithm selection model construction, and model testing processes. In order to eliminate any unnecessary hyper-parameters, the first phase looked at the hyper-parameters for each learning algorithm to assess their respective influence on learning time and model quality. The goal of the second phase was to identify the set of hyper-parameters that created the models with the greatest scores by training models using a variety of hyper-parameters. In order to measure the parallel performance of each algorithm on each dataset, the third phase reran a few selected iterations of the model training processes on varying numbers of compute cores. Phases two and three were joined in the fourth and final stage to create an algorithm selection model.

VI. SNAPSHOTS



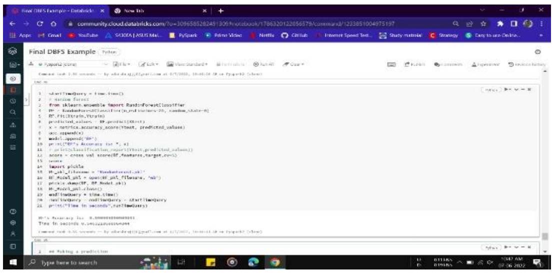
```
from sklearn.model_selection import train_test_split
Xtrain, Xtest, Ytrain, Ytest = train_test_split(features,target,test_size = 0.2,random_state =2)
# Decision Tree
from sklearn.tree import DecisionTreeClassifier
DecisionTree = DecisionTreeClassifier(criterion="entropy",random_state=2,max_depth=5)
DecisionTree.fit(Xtrain,Ytrain)
predicted_values = DecisionTree.predict(Xtest)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('Decision Tree')
print("DecisionTrees's Accuracy is: ", x*100)
# print(classification_report(Ytest,predicted_values))
from sklearn.model_selection import cross_val_score
score = cross_val_score(DecisionTree, features, target,cv=5)
score
import pickle
DT_pkl_filename = 'DecisionTree.pkl'
DT_Model_pkl = open(DT_pkl_filename, 'wb')
pickle.dump(DecisionTree, DT_Model_pkl)
DT_Model_pkl.close()
endTimeQuery = time.time()
runTimeQuery = endTimeQuery - startTimeQuery
print("Time in seconds",runTimeQuery)

DecisionTrees's Accuracy is: [illegible]
Time in seconds [illegible]

startTimeQuery = time.time()
# Gaussian Naive Bayes
from sklearn.naive_bayes import GaussianNB
```

Fig. 6.1 Decision Tree

This image shows the decision tree algorithm's source code, which was used to forecast the accuracy and execution time of the dataset. In actuality, this is regarded as the training phase for determining whether this algorithm is best suited for advising the best crop suitable for high yield in the presence of whether and other characteristic variables.



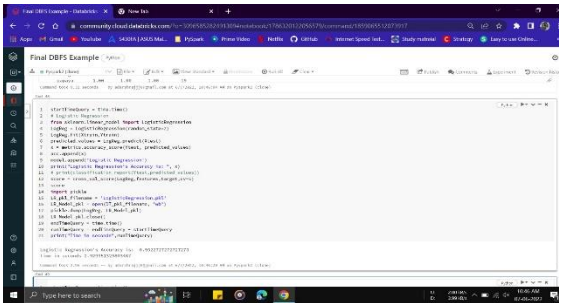
```
startTimeQuery = time.time()
# Random Forest
from sklearn.ensemble import RandomForestClassifier
RF = RandomForestClassifier(n_estimators=20, random_state=0)
RF.fit(Xtrain,Ytrain)
predicted_values = RF.predict(Xtest)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('RF')
print("RF's Accuracy is: ", x)
# print(classification_report(Ytest,predicted_values))
score = cross_val_score(RF,features,target,cv=5)
score
import pickle
RF_pkl_filename = 'RandomForest.pkl'
RF_Model_pkl = open(RF_pkl_filename, 'wb')
pickle.dump(RF, RF_Model_pkl)
RF_Model_pkl.close()
endTimeQuery = time.time()
runTimeQuery = endTimeQuery - startTimeQuery
print("Time in seconds",runTimeQuery)

RF's Accuracy is: [illegible]
Time in seconds [illegible]

## Making a prediction
```

Fig. 6.2 Random forest

This image shows the random forest algorithm's source code, which was used to forecast the accuracy and execution time of the dataset. In actuality, this is regarded as the training phase for determining whether this algorithm is best suited for advising the best crop suitable for high yield in the presence of whether and other characteristic variables. The PySpark API is used to import some of the library sets.

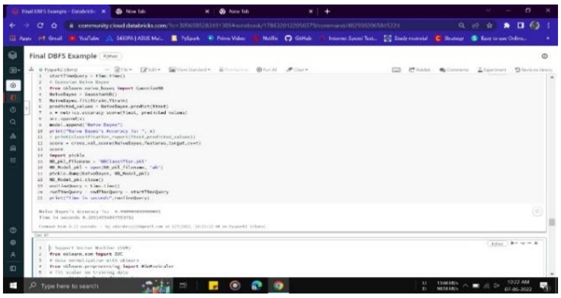


```
startTimeQuery = time.time()
# Logistic Regression
from sklearn.linear_model import LogisticRegression
LogReg = LogisticRegression(random_state=2)
LogReg.fit(Xtrain,Ytrain)
predicted_values = LogReg.predict(Xtest)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('Logistic Regression')
print("Logistic Regression's Accuracy is: ", x)
# print(classification_report(Ytest,predicted_values))
score = cross_val_score(LogReg,features,target,cv=5)
score
import pickle
LR_pkl_filename = 'LogisticRegression.pkl'
LR_Model_pkl = open(DT_pkl_filename, 'wb')
pickle.dump(LogReg, LR_Model_pkl)
LR_Model_pkl.close()
endTimeQuery = time.time()
runTimeQuery = endTimeQuery - startTimeQuery
print("Time in seconds",runTimeQuery)

Logistic Regression's Accuracy is: 0.9522727272727273
Time in seconds 2.923951329803467
```

Fig. 6.3 Logistic Regression

This image shows the source code for the Logistic Regression technique, which was used to forecast the accuracy and execution time of the dataset. In actuality, this is regarded as the training phase for determining whether this algorithm is best suited for advising the best crop suitable for high yield in the presence of whether and other characteristic variables. The PySpark API is used to import some of the library sets.



```
startTimeQuery = time.time()
# Gaussian Naive Bayes
from sklearn.naive_bayes import GaussianNB
NaiveBayes = GaussianNB()
NaiveBayes.fit(Xtrain,Ytrain)
predicted_values = NaiveBayes.predict(Xtest)
x = metrics.accuracy_score(Ytest, predicted_values)
acc.append(x)
model.append('Naive Bayes')
print("Naive Bayes's Accuracy is: ", x)
# print(classification_report(Ytest,predicted_values))
score = cross_val_score(NaiveBayes,features,target,cv=5)
score
import pickle
NB_pkl_filename = 'NBClassifier.pkl'
NB_Model_pkl = open(NB_pkl_filename, 'wb')
pickle.dump(NaiveBayes, NB_Model_pkl)
NB_Model_pkl.close()
endTimeQuery = time.time()
runTimeQuery = endTimeQuery - startTimeQuery
print("Time in seconds",runTimeQuery)

Naive Bayes's Accuracy is: [illegible]
Time in seconds [illegible]

# Support Vector Machine (SVM)
from sklearn.svm import SVC
# data normalization with sklearn
from sklearn.preprocessing import MinMaxScaler
# fit scaler on training data
```

Fig 6.4 Naïve bayes

This image displays the source code which contains the Naïve Bayes algorithm used for predicting the accuracy and time taken for executing the dataset. This is actually considered as the training phase for knowing whether this algorithm is best suited for recommending the best crop suitable for high yield in provided whether and other characteristic conditions. Some of the library sets are imported from PySpark API.

VII. CONCLUSION

Agriculture depends on being able to predict the best crop to grow. In order to determine which crops would be the best to cultivate, experiments were done utilising the NB, DT, Logistic Regression, RF, and Map Reduce methodologies. For an investigation of the crop prediction process, soil and environmental parameters were taken into account. Using multiple machine learning techniques, the findings show varying accuracy. We are using Spark cluster to deploy and lower the time limitation because of the time factor for doing this project. We also came to the conclusion that the cluster processes large datasets more quickly and can provide predictions. When the location and climatic conditions are taken into consideration, the top 20 crops that address the bulk of repercussions are used in this paper proposal. Farmers will be able to choose an appropriate crop for yield based on how accurately various crops are predicted around the world.

VIII. REFERENCES

[1] Zeel Doshi,Subhash Nadkarni ,Rashi Agrawal &Prof. Neepa Shah ,"AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms"

[2] M.V.R. Vivek, D.V.V.S.S. Sri Harsha, P. Sardar Maran, "A Survey on Crop Recommendation Using Machine Learning",International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5C, February 2019

[3] A. Suruliandi, G. Mariammal & S.P. Raja (2021) Crop prediction based on soiland environmental characteristics using feature selection techniques, Mathematical and Computer Modelling of Dynamical Systems, 27:1, 117-140, DOI: 10.1080/13873954.2021.1882505

[4]Vaishnavi.S1, Shobana.M 2.Sabitha.R3, K arthik.S,"Agricultural Crop Recommendations based on Productivity and Season" 7th International Conference on Advanced Computing and Communication Systems | 978-1-6654-0521-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICACCS51430.2021.9441736

[5] G. Mariammal , A. Suruliandi , S. P. Raja , and E. Poongothai, "Prediction of Land Suitability for Crop Cultivation Based on Soil and Environmental Characteristics Using Modified Recursive Feature Elimination Technique With Various Classifiers"

[6] Shiyam Talukder,Habiba Jannat,Katha Sengupta,Sukanta Saha and Muhammad Iqbal Hossain ,"Enhancing Crops

Production Based on Environmental Status Using Machine Learning Techniques ",Authorized licensed use limited to: University of Technology Sydney.

[7] Mummaleti Keerthana,K J M Meghana ,Siginamsetty Pravallika and Dr. Modepalli Kavitha, "An Ensemble Algorithm for Crop Yield Prediction " Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021).
IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4|DOI: 10.1109/ICICV50876.2021.9388479

[8] M. A. Hall and G. Holmes, "Benchmarking clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[9] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," *Chemometric Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, Sep. 2006.

[10] Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Jan. 2003

[11] P. S. Maya Gopal and R. Bhargavi, "Feature selection for yield prediction in boruta algorithm," Int. J. Pure Appl. Math., vol. 118, no. 22, pp. 139–144, 2018.