Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Prediction of Crop Yield Using Efficient Data Analysis Techniques

Y Pavan Venkata Sai[1], V Naveen Kumar Redddy[2], T Srihari[3], V. Ramesh Babu[4], R. Selvameena[5,] M. Anand[6]

[4,6]Professor,[5] Asst.Professor,[1,2,3]UG Students,
[1,2,3,4,5,6]Department of Computer Science and Engineering,
[1,2,3,4,5,6]Dr. M. G. R Educational and Research Institute of Technology, Madhuravoyal, Chennai, Tamil Nadu, India.
pavanyekula1234@gmail.com, rameshbabu.cse@drmgrdu.ac.in

Abstract

The weather's impact on crop output might be regarded as a crucial factor in agricultural yield forecast. Numerous studies have been undertaken to determine how weather impacts agriculture, however the majority of these studies need vast amounts of complicated data that are not readily accessible. As a result, the approach must be improved to account for the lack of data. Machine learning (ML) can extract patterns and associations from datasets and derive information from them. We gathered agricultural and meteorological data from a vast dataset to ensure reliable crop forecasting. We chose Random forest, SVM, and Decision tree as the best models for the proposed system since they are more efficient and perform better than current models at calculating extremely big datasets of weather and climate, assuring greater accuracy and faster processing. We provide input characteristics such as weather, rainfall, and temperature and obtain the most ideal crop as a result. The primary objective of this model is to forecast changes in the weather and to assist farmers in making agricultural choices in response to such changes. Additionally, data mining is beneficial for forecasting agricultural yield output. In general, data mining is the act of examining data from several perspectives and condensing it into useful information.

Random forest is the most widely used and powerful supervised machine learning algorithm. Random forest is the most widely used and powerful supervised machine learning algorithm.

Keywords: Crop Yield Prediction, Data Analysis Techniques, Agricultural yield forecast, Machine Learning, Random forest, Support Vector Machine, Decision tree, Data Mining.

1. INTRODUCTION

Nowadays, many people are unaware of the importance of cultivating crops at the proper time and location. Seasonal climatic conditions are also altered as a result of these cultivating practices, putting key assets like as land, water, and air at risk, resulting in food insecurity

Climate change has had a detrimental effect on the performance of the majority of agricultural crops in India during the previous two decades. Predicting crop yields in advance of harvest enables policymakers and farmers to take proper marketing and storage strategies

This initiative will assist farmers in determining the yield of their crop prior to growing it on the agricultural field, enabling them to make informed choices. It is based on the random forest algorithm.

It makes an effort to resolve the problem by developing an interactive prediction system prototype. The system will be implemented using an easy-to-use web-based graphical user interface and the machine learning algorithm. The farmer will be informed of the outcome of the forecast. Thus, for this kind of data analytics in crop prediction, several methodologies or algorithms exist, and we may forecast crop production using those algorithm

It performs classification and regression tasks by training a large number of decision trees and generating output for the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Appropriate guidance and analyses of future agricultural yield are required to assist farmers in maximizing crop output. Predicting yields is a significant agricultural concern. There are several methods for increasing and improving agricultural output and quality. Additionally, data mining is beneficial for forecasting agricultural yield output.

2. RELATED WORK

[1] Wang.A,et al.,2022 has represented an idea to collect and analyze data on temperature, precipitation, soil, seeds, crop production, humidity, and wind speed. First, preprocess the data in the Python environment before further analyzing and processing large amounts of data using the MapReduce framework. Then use K-means clustering for MapReduce results to get meaningful results for your data in terms of accuracy. In addition, we used a self-designed recommender system to predict yieldsand display them in a Flask-based graphical user interface. We all agree that agriculture in India is the backbone of the country.

[2] Potnuru Sai Nishant,et al.,2020 has represents predicts yields for almost all plant species grown in India. This script is new because it uses basic criteria such as state, district, season, region, etc., and allows users to predict crop yields for any year. This study uses advanced regression techniques such as the Kernel Ridge, Lasso, and ENet algorithms, and the concept of stacking regression to improve the algorithm to predict yield.

[3] Y.JeevanNagendra,,et al.,2020 has represented the idea get the output you want,you need to generate the appropriate function with a collection of variables that mapthe input variables to the desired output. Yield predictions include predicting yieldsbased on historical data that includes factors such as temperature, humidity, pH, rainfall, and plant names. It provides us with the best crop ideas that can be produced inthe field under given weather conditions. Random Forest, a machine learning system,can make these predictions. The most accurate crop predictions are achieved. Therandom forest method is used to find the best yield model by examining the smallestnumber of models. Use data mining and machine learning algorithms to extract important information that can help you predict or recommend the best crops.

[4] Anupama ,et al.,2019 proposed algorithm is good at accurately estimating crops, parameters used, different sources used, and features used, all in common with accuracy and error rate. There are many possibilities for growth and expansion. Feasibility analysis of smart farming relies heavily on cloud-based big data analytics andIoT technologies. Now, we will use IoT Ice.

[5] S.Rajeswari ,et al.,2017. This author explains the data is used to collect agricultural data and store it in a cloud database. Cloud-based big data analysis is used to evaluate data such as fertilizer needs, crop analysis, crop market, and storage needs.

[6]Aruvansh Nigam,et al.,2019. This author focus on predicting yields by applying a variety of machine learning techniques. The results of these techniques are compared based on mean absolute error. Machine learning algorithm predictions help farmers decide which crop to plant for maximum yield, taking into account factors such as temperature, rainfall, and area.

[7]R. Ghadge,et al.,2018. This author focused on machine learning-based frame work for crop yield prediction. A dataset of attachment details is created for the setup ofthe experiment. The investigation uses the machine learning algorithms SVM, Random Forest, and ID3.

[8] F. H. Tseng,et al.,2018. This author research to the helps the beginner farmers in such a way to guide them in sowing reasonable crops by deploying machine learning, one of the advanced technologies in crop prediction. Naive Bayes, a supervised learning algorithm puts forth the way to achieve it. The seed data of the crops are collected here, with the appropriate parameters like temperature, humidity, and moisture content, which helps the crops to achieve successful growth.

[9] Suresh, N. Manjunathan,et al.,2020. These authors implementes the machine learning techniques along with the remote sensing data to extract features. Support Vector Machine and KNearest Neighbor techniques that have shown outstanding results are implemented to predict the crop yield using remote sensing data.

[10] P. Sivanandhini,et al.,2020. This author describes about Random Forest, the most common and powerful supervised machine learning algorithm that can handle both classification and regression tasks. They are used in crop selection to reducecrop yield loss, regardless of the distracting environment. The effects of weather, climate, and other related environments pose a significant threat to the long-term survival of agriculture.

3. METHODOLOGY

3.1 General Architecture

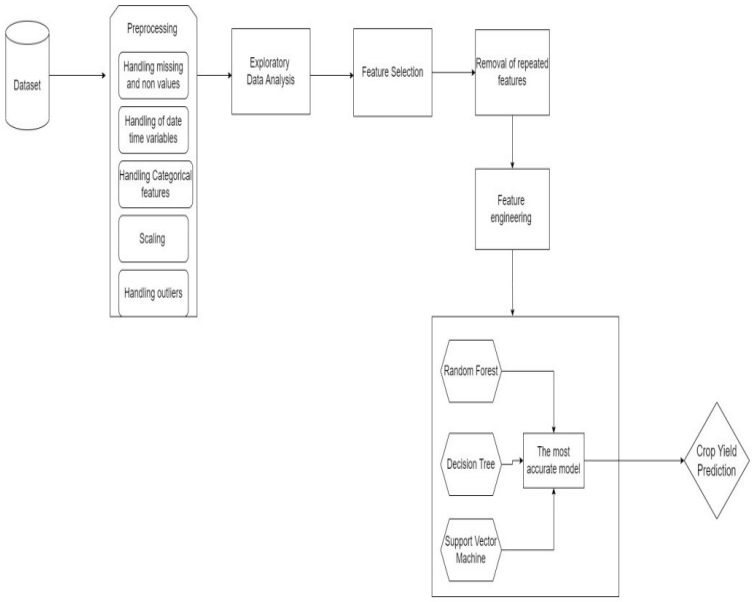

Figure 4.1: **System Architecture of Crop Yield Prediction**

Fig 4.1 explains to collect data set after that data set has to be Preprocessing .In that preprocessing the data can handle the missing and non values, Handling of date variables ,Handling the categorical features and scaling .After Preprocessing of the dataset we can use EDA Exploratory Data Analysis .EDA is used to analyze and investigate datasets and summarize their main characteristics. After this EDA can do feature selection for the given data set in that we can remove the repeated features. After removing all the repeated features we can find which is the most accurate model for the crop yield Prediction. This project contains Random forest ,Decision Tree and Support Vector Machine to find the accuracy.

3.2 Design Phase

3.2.1 Data Flow Diagram

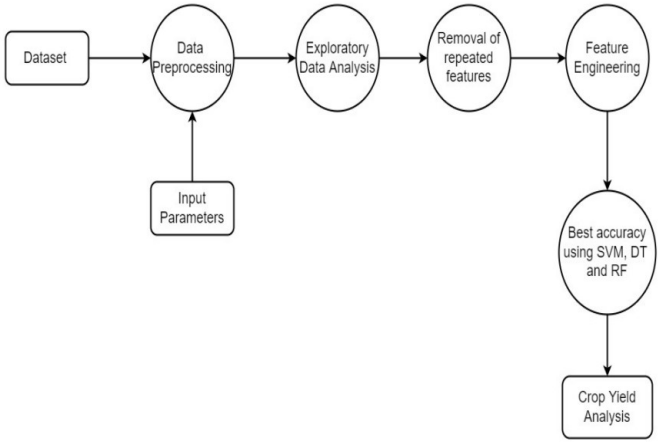

Figure 4.2: **DFD diagram of Crop Yield Prediction**

Fig 4.2.1 describes about the data set has to be Preprocessing. Before Preprocessing the dataset we have to give input parameters .In that preprocessing we can handle the missing and non values, Handling of date variables, Handling the categorical features and scaling .After Preprocessing of the dataset we can use EDA Exploratory Data Analysis .EDA is used to analyze and investigate datasets and summarize their main characteristics. After this EDA we can do feature selection for the given data set in that we can remove the repeated features. After removing all the repeated features we can find which is the most accurate model for the crop yield Prediction. We use Random forest ,Decision Tree and Support Vector Machine to find the accuracy.

3.2.3 Class Diagram

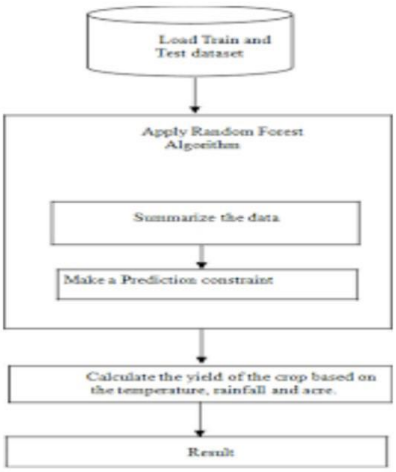

Figure 4.4: **Class Diagram of Crop Yield Prediction**

Fig 4.2.3 describes about the data set has to be Preprocessing. Before Preprocessing the dataset we have to give input parameters .In that preprocessing we can handle the missing and non values, Handling of date variables, Handling the categorical features and scaling .After Preprocessing of the dataset we can use EDA Exploratory Data Analysis .EDA is used to analyze and investigate datasets and summarize their main characteristics. After this EDA we can do feature selection for the given data set in that we can remove the repeated features. After removing all the repeated features we can find which is the most accurate model for the crop yield Prediction. We use Random forest ,Decision Tree and Support Vector Machine to find the accuracy.

3.2.4 Sequence Diagram

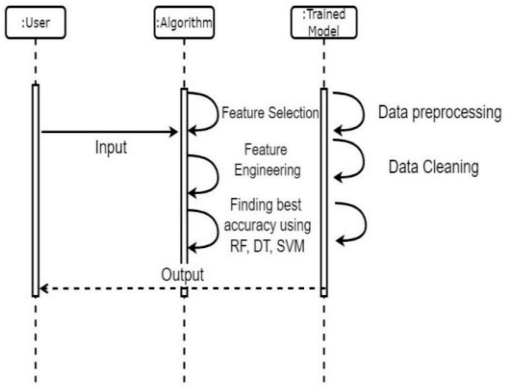

Figure 4.5: **Sequence Diagram of Crop Yield Prediction**

Fig 4.2.4 shows that the user has to give the input after that the data set can be preprocessing and cleaned. After these process we have to find out the feature selection and we have to find the accuracy using Random Forest, Decision Tree and Support vector Machine. Using that three classifiers we can tell the user which crop is suitable for that weather conditions.

3.2.5 Collaboration diagram

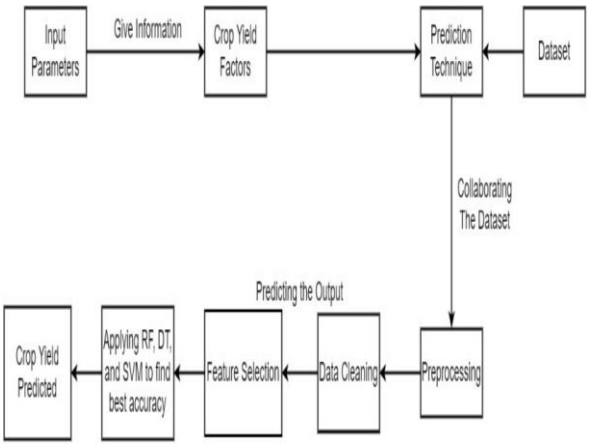

Figure 4.6: **Collaboration Diagram of Crop Yield Prediction**

Fig 4.2.5 represent the interactions between objects and their behaviors in the context to a particular use case. In simple terms, these diagrams clarify the roles played by different objects. This explains about the data set has to be Preprocessing. Before Preprocessing the dataset we have to give input parameters. In that preprocessing we can handle the missing and non values, Handling of date variables, Handling the categorical features and scaling .After Preprocessing of the dataset we can use EDA Exploratory Data Analysis .EDA is used to analyze and investigate datasets and summarize their main characteristics. After this EDA we can do feature selection for the given data set in that we can remove the repeated features. After removing all the repeated features we can find which is the most accurate model for the crop yield Prediction. We use Random forest ,Decision Tree and Support Vector Machine to find the accuracy.

3.3 Module Description

3.3.1 Identifying the Statement

Determine the problem statement which concise description of a problem or a project seeks to address.

3.3.2 Collection of dataset

Data Collection and training using Machine Learning Algorithms and making the

dataset ready to work on.

3.3.3 Conceptualize the data

Conceptualize our data using scatter plot, distribution graph, etc. by doing so we can find out anomalies, missing values, etc. on our data and make our dataset perfect for

prediction

3.3.4 Selection of dataset

Firstly we have searched data sets from various resources. Later we found a better data set in github. After collecting that data set we preprocessed that dataset .In preprocessing we handle the missing and non values, Handling of date variables ,Handling the categorical features and scaling

3.3.5 Applying classifiers

After importing that data we choose three machine learning classifiers like Decision

Tree,Random Forest, Support Vector Machine then we compare all these classifiers

for better prediction.Now we have to check which algorithm will have high accuracy. we used another algorithm but it does not give better accuracy than the above

algorithms.

3.3.6 Feature Selection

The feature removal procedure begins after the preprocessing of given data and EDA

(Exploratory data analysis). When creating a predictive model, the technique for feature selection refers to the process of limiting the number of inputs. It is a procedure in which repetitive characteristics from the pre-processed dataset are deleted and only the most significant ones are retained in order to develop the best weather based crop yield prediction model. Certain predictive modelling challenges include a huge number of variables, which may significantly slow down model construction and training and consume a significant amount of system memory. Reduce the number of input variables to both lower the computational cost of modelling and, in certain situations, increase the model's performance. We tested many models fitted to various subsets of characteristics selected using various statistical metrics and observed that the suggested model performed the best in the area of crop prediction.

Feature selection can be done in a lot of ways using machine learning.

3.3.7 Feature Importance:

Decision trees and the random forest is used to approximate the vitality of features. Attributes that are scored higher are of more importance than others.

3.3.8 Building the best accuracy producing algorithm:

This module uses a wide range of algorithms. It commences the process of model development and evaluation. Random Forest, Decision Tree, and Support Vector Machine are among the methods employed. The Random Forest Algorithm generates the final result by combining the output of many Decision Trees. We picked the decision tree because it is very simple to grasp, even for those with no analytical expertise. The SVM algorithm's objective is to find the optimal line or decision boundary that partitions n-dimensional spaces into classes, allowing us to easily classify fresh data points in the future. We will save the method that generates the best

result and use it as the output. We provide input characteristics such as weather, rainfall, and temperature and

obtain the most ideal crop as a result.

4 Algorithm & Pseudo Code

4.1 Algorithm

1.Data Pre-processing step.

2.Applying the EDA to that data set.EDA is used to analyze and investigate data sets and summarize their main characteristics.

3.Feature selection

4.Removal of repeated features.

5.Apply the feature engineering.

6.Finding the accurate model in RD,SVM and DT.

7. Predict which crop is suitable for yield at that weather conditions.

4.2 Pseudo Code

```
Feature Scaling
from sklearn.preprocessing import Standard Scaler
scX = StandardScaler()
Xtrain = scX.f ittransform(Xtrain)
Xtest = scX.transform(Xtest)
Fitting Random Forest to data set
from sklearn.linearmodel
import Random Forest
classif ier = Random Forest ()
classif ier.f it(Xtrain, ytrain)
Predicting the test set result
ypred = classif ier.predict(Xtest)
Making the confusion matrix
from sklearn.metrics import confusionmatrix
cm = confusionmatrix(ytest, ypred
```

5 FUTURE SCOPE

This approach uses many algorithms to predict yield reliably and efficiently. The general goal of this project is to analyze data from data collections and use this data to make forecasts of agricultural production. Not 100 percent accurate, but this application is much more accurate than other methods. In the future, it can be expanded to suggest fertilizers, appropriate guidelines for cultivable land, and crops for specific inputs. In addition, solar radiation and plant health are regularly monitored and used to increase crop yields.

6. CONCLUSION

Agriculture is the backbone of the Indian economy, this application can help farmers make informed decisions about crop planting in a variety of ways. To achieve this, the data is provided through Flask web pages and is supported by powerful machine learning methodologies and technologies, and an intuitive user experience.

Agricultural prices can be predicted using a number of methods such as decision trees, support vector machine,Random Forest.

7. ACKNOWLEDGEMENT

We express our sincere gratitude to Vishwakarma Institute of Technology for providing the platform to develop this remarkable project. Special thanks to our esteemed Director, Prof. Dr. Rajesh M. Jalnekar, for his invaluable guidance and unwavering support. We also acknowledge the support from Prof. (DR.) Sandeep Shinde, Head of the Department of Computer Engineering, and extend our appreciation to our project guide, Prof. Rakhi Bhardwaj, for her valuable time, support, and inspiration in creating this paper.

REFERENCES

[1] Wang, A., Chen, G., Yang, J., Zhao, S., Chang, C. Y. (2016).“A comparativestudy on human activity recognition using inertial sensors in a smartphone.”IEEE Sensors Journal, 16(11), 4566-4578.

[2] Potnuru Sai Nishant, Pinapa Sai Venkat, Bollu Lakshmi Avinash, and B. Jabber, 2020, “Crop Yield Prediction based on Indian Agriculture using MachineLearning,” 2020 International Conference for Emerging Technology (INCET).

[3] Y. Jeevan Nagendra Kumar, V. Spandana, V.S. Vaishnavi, K. Neha, V.G.R.R.Devi, 2020, “Supervised Machine learning Approach for Crop Yield Predictionin Agriculture Sector,” 2020 5th International Conference on Communicationand Electronics Systems (ICCES).

[4] Anupama C.G., Lakshmi C, 2019, “A comprehensive review on the crop prediction algorithms,” Next Generation Computing Technologies (NGCT), 2019 1st International Conference.

[5] S. Rajeswari, K. Suthendran, K. Rajakumar, 2017, “A smart agricultural modelby integrating IoT, mobile and cloud-based big data analytics,” 2017 International Conference on Intelligent Computing and Control (I2C2).

[6] Aruvansh Nigam; Saksham Garg; Archit Agrawal; Parul Agrawal, 2019, “CropYield Prediction Using Machine Learning Algorithms”, 2019 IEEE Fifth International Conference on Image Information Processing (ICIIP)

[7] R. Ghadge, J. Kulkarni, P. More, S. Nene and R. L. Priya, “Prediction of cropyield using machine learning”, Int. Res. J. Eng. Technology, vol. 5, 2018.

[8] F. H. Tseng, H. H. Cho and H. T. Wu, “Applying big data for intelligentagriculture-based crop selection analysis”, IEEE Access, vol. 7, pp. 116965-116974, 2019.

[9] A. Suresh, N. Manjunathan, P. Rajesh and E. Thangadurai, “Crop Yield Prediction Using Linear Support Vector Machine”, European Journal of MolecularClinical Medicine, vol. 7, no. 6, pp. 2189-2195, 2020.

[10] P. Sivanandhini and J. Prakash, “Crop Yield Prediction Analysis using FeedForward and Recurrent Neural Network”, International Journal of InnovativeScience and Research Technology, vol. 5, no. 5, pp. 1092-1096, 2020.

[11] S. D. Kumar, S. Esakkirajan, S. Bama and B. Keerthiveena, “A microcontrollerbased machine vision approach for tomato grading and sorting using SVM classifier”, Microprocessors and Microsystems, vol. 76, pp. 103090, 2020.

[12] B. Devika and B. Ananthi, “Analysis of crop yield prediction using data miningtechnique to predict annual yield of major crops”, International Research Journalof Engineering and Technology, vol. 5, no. 12, pp. 1460-1465, 2018.

[13] V. Pandith, H. Kour, S. Singh, J. Manhas and V. Sharma, “Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction fromSoil Analysis”, Journal of Scientific Research, vol. 64, no. 2, 2020.

[14] D. A. Bondre and S. Mahagaonkar, “Prediction of Crop Yield and FertilizerRecommendation Using Machine Learning Algorithms”, International Journalof Engineering Applied Sciences and Technology, vol. 4, no. 5, pp. 371-376,2019.

[15] Sekhar Sajja; Subhesh Saurabh Jha; Hicham Mhamdi; Mohd Naved; SamratRa, 2021, “An Investigation on Crop Yield Prediction Using Machine Learning”,2021 IEEE Third International Conference on Inventive Research in ComputingApplications (ICIRCA).