Crop Yield Prediction Using Machine Learning Approaches

Dr.M.Sengaliappan[1], Bharathkumar.K[2]
[1]Head of the Department, Department of Computer Applications,
Nehru college of management, Coimbatore, Tamilnadu, India
E-mail-ncmdrsengaliappan@nehrucolleges.com
[2]Student of II MCA, Department of Computer Applications,
Nehru College of Management, Coimbatore, Tamilnadu, India
E-mail-bk3006848@gmail.com

Abstract
Farming accounts for the majority of India's GDP, which is reliant on agriculture. This project is motivated by the rising suicide rates among farmers, which could be brought on by poor agricultural yields. The agricultural industry is now seriously threatened by changes in the climate and other environmental factors. To find workable and efficient solutions for this issue, machine learning is a crucial strategy. using past data, such as weather, soil, rainfall, and previous crop yield, to forecast crop yield. The machine learning algorithm helped us do this. KNN- K Nearest Neighbour, Random Forest, Decision tree, and Logistic Regression were among the machine learning methods we compared. The Random Forest approach yielded an accuracy rate. A web application that forecasts crop yields both generally and for a specific crop has been developed as part of this research. It also tells the user whether or not the time is ideal to apply the fertilizer.

Keywords-KNN, Decision tree, Random Forest, logistic regression, python programming; confusion matrix; correlation matrix

1.INTRODUCTION

With a population of over 1.3 billion, India is the second-largest country in the world. The agricultural sector, particularly in our nation, lacks efficiency and technology, despite the fact that many people rely on it. Effective crop cultivation can be accomplished by establishing a connection between data science and conventional agriculture. For many people in developing nations, farming is their primary source of income. These days, a variety of inventions, settings, methods, and civilizations contribute to the expansion of agriculture. Additionally, the use of information technology may alter the way that decisions are made, allowing farmers to produce the best results. Agricultural data mining techniques are employed for the decision-making process. Analysing, extracting, and forecasting significant information from massive amounts of data in order to identify patterns is known as data mining. This procedure is utilized by corporations to change the raw data of their consumer to usable information. The initial step in the data mining process is data selection, which is followed by pre-processing and data transformation to identify patterns that can be utilized to forecast insightful information. While preprocessing involves identifying missing values and outliers, transformation determines the relationship between items. A number of forecasts can be produced based on the information acquired by using data mining techniques on past climate and crop production data, which can ultimately aid in raising crop yield.

It is possible to retrieve the most recent rainfall and climate data. Each of the values is posted in the direction of R. In order to forecast the climate changes from the previous year, the collected data is examined using the agriculture dataset. This will provide the farmer with a way to cultivate the crop that will yield the highest profits. With KNN, SVM, and random forest, variations in temperature, humidity, rainfall, and pH can be identified. Businesses may make proactive, informed decisions by using data mining techniques to forecast future trends and behaviours. Beyond the analysis of past events provided by display tools commonly found in decision support systems, data mining offers motorized, prospective analyses. Business problems that once took too much time to address can now be answered with data mining techniques. The district is irrigated based on temperature and rainfall. The initial set is the root node (rainfall) at the start of the process. Each non-terminal node (internal node: temperature range) in the decision tree represents the chosen attribute that was used to split the data, and the terminal nodes (leaf nodes: district) provide the class label of the last subset of this branch.

2.LITERATURE SURVEY

[1] "A Method to Build a Decision Tree with Sliq and Knn for Land Grading System," by K.K. Joshi and P. Patidar, ISSN: 2277-5528, Impact Factor: 2.745 (Sijf), pp. 231-236.

Patidar P. and Joshi K. K. demonstrated an approach to agricultural data mining. Only pertinent attributes were chosen for additional use once feature selection procedures were used. Several feature selection methods were employed in the Quest (Sliq) Algorithm, including ID3 Decision Tree Basics, Entropy, Information Gain, and Supervised Learning. The findings of their comparison of those feature selection methods are displayed.

[2] A Survey on Decision Tree Algorithm for Classification by Brijain R. Patel and Kushik K. Rana, International Journal of Engineering Development and Research, Vol. 2, Issue 1, 2014.

According to Brijain R. Patel and Kushik K. Rana, classification and prediction are methods used to identify significant data classes and forecast likely trends.In data mining classification, the Decision Tree is a crucial classification technique. It is frequently employed in scientific research, marketing, monitoring, and fraud detection.

[3] "Internet of Things Based System For Remote Monitoring of Weather Parameters and Applications," International Journal of Advances in Electronics and Computer Science, Vol. 3, Issue 2, February 2016; Prachi H. Kulkarni and Pratik D. Kute.

Prachi H. Kulkarni discussed the proof of concept for an Internet of Things device that uses a complex microcontroller platform to gather physical parameter data from multiple sensor kinds over numerous communication channels before uploading the data to the Internet. The equipment that is being demonstrated is intended for remote weather parameter monitoring.

3.MACHINE LEARNING APPROCHES

We have employed four popular machine learning techniques to create the crop yield prediction model. The details of these tactics are as follows:

3.1 Logistic Regression:

A classification approach for supervised learning, logistic regression is used to forecast the likelihood of the target variable. There would only be two viable classes because the goal or dependent variable is dichotomous. Upon applying the logistic regression algorithm to our dataset can be applied.

3.2 Decision Tree:

The Decision Tree algorithm is classified as a supervised learning technique. The decision tree approach can be used to forecast regression and classification issues, whereas the supervised learning technique provides a solution at an accurate prediction level. The Decision Tree method can be used to anticipate crops based on temperature ranges and climatic variations. Decision trees can be used to build the training model for crop prediction. The class's target variables can be determined by studying the decision rules that were drawn from previous information (training data). State, district, crop year, season, yearly rainfall, minimum and maximum temperatures, and crop are among the fields included in the training data. The decision tree algorithm is used to build the tree model. The leaf node corresponds to a class name, while each internal node represents a different property of the data set.

3.3 Random Forest Classification:

A Crop development in relation to current climatic circumstances and biophysical change can be analysed by Random Forest. The Random Forest algorithm builds decision trees using several data samples, forecasts the data from each subset, and then uses voting to provide a better system solution. The bagging technique is used by Random Forest to train the data, improving the result's accuracy.

3.4 K-Nearest Neighbor:

A K-Nearest Neighbor (KNN) can be applied to crop yield prediction in both regression and classification settings. KNN is a non-parametric and simple algorithm that can predict crop yield based on similarities between past data points and current input conditions (like weather, soil type, irrigation, etc.).

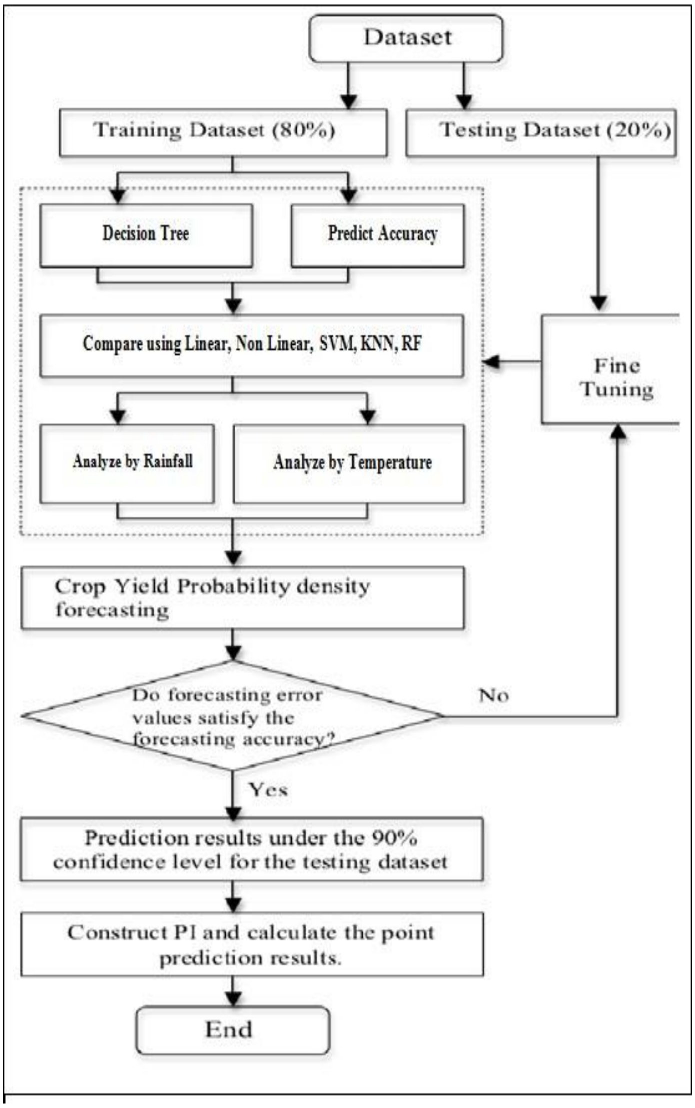
3.5 Dataset

- https://ourworldindata.org/crop-yields
- https://www.kaggle.com/datasets/thedevastator/the-relationship-between-crop-production-and-cli

This dataset offers information on wheat, maize, rice, and soybean production amounts, harvested areas, and crop yields. Crop yields are the amount of crop goods produced per unit of harvested area. Yield data are typically not recorded; instead, they are calculated by dividing production data by area harvested data. The genetic potential of the crop, the amount of sunlight, water, and nutrients it absorbs, as well as the presence of weeds and pests, all affect the actual yield that is harvested on a farm. This indication is offered for wheat, maize, rice, and soybean. Tonnes per hectare are used to measure crop yield.

4. METHODOLOGY

The process used to construct the crop yield prediction model is shown in the steps below.


Dataset
Training Dataset (80%)
Testing Dataset (20%)
Decision Tree
Predict Accuracy
Compare using Linear, Non Linear, SVM, KNN, RF
Fine Tuning
Analyze by Rainfall
Analyze by Temperature
Crop Yield Probability density forecasting
Do forecasting error values satisfy the forecasting accuracy?
No
Yes
Prediction results under the 90% confidence level for the testing dataset
Construct PI and calculate the point prediction results.
End

4.1 DATA COLLECTION

We used the Crop yield Dataset, which is available online at the UCI Repository [16]. The 14 attributes taken into account are as follows:

Table.1.Attributes

S.no	Attribute	Description
01	State_Name	Region name
02	Crop_year	Harvest year
03	Area	Land area
04	Yield	Crop production
05	Temperature	Average temperature
06	Precipitation	Rainfall amount
07	Humidity	Air moisture
08	Soil_type:_chalky	Chalky soil
09	Soil_type:_clay	Clay soil
10	Soil_type:_loamy	Loamy soil
11	Soil_type:_peaty	Peaty soil
12	Soil_type:_sandy	Sandy soil
13	Soil_type:_silt	Silt soil
14	Soil_type:_silty	Silty soil

The dataset for crop yield prediction includes attributes such as State_Name, which identifies the region or state where the crops are grown, and Crop_Year, representing the year of harvest. Area refers to the size of the land used for crop cultivation, and Yield indicates the crop yield per unit of area. Temperature, Precipitation, and Humidity are climatic factors that influence crop growth, while various Soil_Type attributes, such as chalky, clay, loamy, peaty, sandy, silt, and silty, describe the soil conditions in which the crops are planted. These factors, when combined, provide a comprehensive understanding of the conditions that affect crop yield in a given region.

4.2 Data Preprocessing:

4.2.1 Missing Values:

During preprocessing, the dataset's potential missing values must be appropriately handled (e.g., through imputation, eliminating rows with missing values, etc.). For instance, you might use the average yield from other years or areas to fill in the gaps in agricultural production data for certain years or regions, or you may use alternative approaches like interpolation.

4.2.2 Feature Scaling:

To make sure that every feature contributes equally to model performance, scaling may be required if there are numerical attributes (such as temperature, precipitation, and soil type).

4.2.3 Class Imbalance:

Techniques such as under sampling the dominant class or oversampling the minority class can be used to balance the dataset if there is an uneven distribution between high and low yield.

4.3 Feature Selection:

Logistic Regression is used for classification tasks where crop yield is categorized (e.g., high or low). It assumes a linear relationship between the input features and the target variable. It's computationally efficient but may not capture non-linear patterns effectively.

4.3.1 K-Nearest Neighbor:

KNN is a simple, non-parametric algorithm that classifies or regresses based on the proximity of data points. It works well for smaller datasets but can be computationally expensive for large ones. It is sensitive to the choice of K and the distance metric.

4.3.2 Random Forest:

Random Forest is an ensemble of decision trees that combines predictions from multiple trees to improve accuracy and reduce overfitting. It handles both classification and regression tasks effectively. It also provides feature importance, helping identify critical variables for crop yield prediction.

4.3.3 Decision Tree:

Decision Tree models are tree-based algorithms that split data based on feature values to predict outcomes. They are highly interpretable and can model complex, non-linear relationships. However, they may overfit without proper tuning or pruning.

5.RESULTS:

At the conclusion of our research, the results of Logistic regression model show more accuracy than the KNN, Decision tree and random forest. In comparison with these models, Logistic regression gives 6% more than the KNN, Decision tree and random forest. Logistic Regression performs the best in terms of precision, recall, F1 score, and accuracy for both classes. It strikes a good balance between identifying true positives and minimizing false positives/negatives.

Table.2.Sample Dataset – 1

Crop	State	Cost of Cultivation (`/Hectare)	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)	Support price
GRAM	Rajasthan	8552.69	12610.85	1691.66	6.83	5100
GRAM	Madhya Pradesh	9803.89	16873.17	1551.94	10.29	5100
GRAM	Uttar Pradesh	12833.04	21618.43	1882.68	10.93	5100
GRAM	Maharashtra	12985.95	18679.33	2277.68	8.05	5100

GRAM	Andhra Pradesh	14421.98	26762.09	1559.04	16.69	5100
GROUNDNUT	Karnataka	13647.1	17314.2	3484.01	4.71	5275
GROUNDNUT	Andhra Pradesh	21229.01	30434.61	2554.91	11.97	5275
GROUNDNUT	Tamil Nadu	22507.86	30393.66	2358	11.98	5275
GROUNDNUT	Gujarat	22951.28	30114.45	1918.92	13.45	5275
GROUNDNUT	Maharashtra	26078.66	32683.46	3207.35	9.33	5275
MAIZE	Bihar	13513.92	19857.7	404.43	42.95	1850
MAIZE	Karnataka	13792.85	20671.54	581.69	31.1	1850
MAIZE	Rajasthan	14421.46	19810.29	658.77	23.56	1850
MAIZE	Uttar Pradesh	15635.43	21045.11	1387.36	13.7	1850

Table.3.Sample Dataset – 2

State Name	Crop Year	Area	Yield	Temperature	Precipitation
Maharashtra	1998	21100	25500	23.9526	15.437
Maharashtra	1999	24200	28800	23.6212	26.8132
Maharashtra	2000	17100	18900	24.478	25.439
Maharashtra	2001	17100	19600	24.7904	6.8608
Maharashtra	2002	17600	18600	24.563	25.609
Maharashtra	2003	17600	20800	23.7682	5.4152
Maharashtra	2004	5100	5000	23.9526	8.2766
Maharashtra	2005	15900	15800	23.6212	32.9788
Maharashtra	2006	33100	37600	24.478	28.729
Maharashtra	2008	38000	41200	24.7904	32.4276
Maharashtra	2009	18000	20000	24.563	40.5846
Maharashtra	2010	41700	61400	24.58152	27.4774
Maharashtra	2011	34100	58500	24.682	23.9894
Maharashtra	2012	36100	64600	24.789	23.9894
Maharashtra	2013	34500	30700	24.854	3.593
Maharashtra	2014	29100	26800	24.647	17.4454

Table.4.Crop Distribution in Percentage

Gram	10.2
Cotton	10.2
Arhar	10.2
Wheat	8.2
Sugarcane	10.2
Rapeseed and mustard	10.2
Paddy	10.2
Moong	10.2
Maize	10.2
GroundNut	10.2

Table.5.Crop Distribution in Percentage

Gujarat	8.2
Karnataka	10.2
Uttar Pradesh	14.3
West Bengal	2
Orissa	4.1
Bihar	2
Tamil Nadu	4.1
Madhya Pradesh	6.1
Rajasthan	10.2
Haryana	4.1
Punjab	6.1
Maharastra	12.2
Andhra Pradesh	16.3

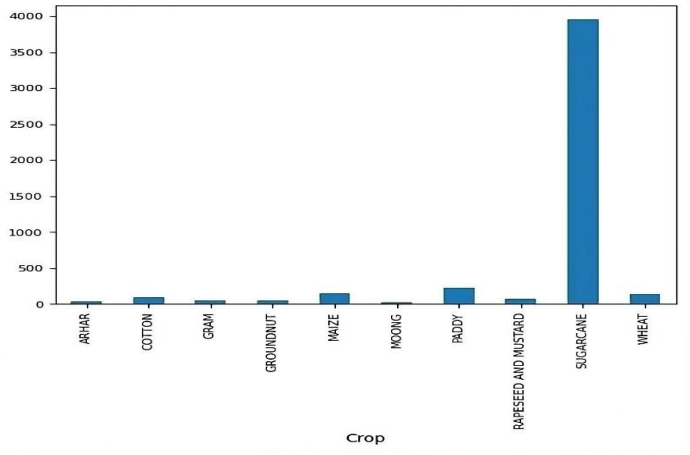


Fig.1.Crop Wise Yield

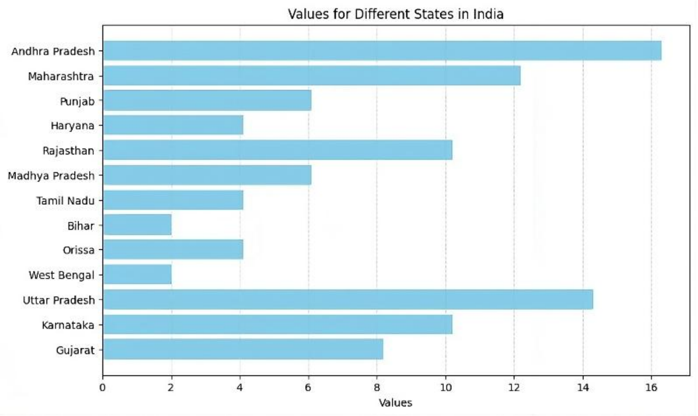


Fig.2.Values in States

Table.6.Overall Results

Algorithm	Precision		Recall		F1 Score		Accuracy
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	
Logistic Regression	1	0.89	0.86	1	0.92	0.94	0.93
Decision Tree	0.78	1	1	0.74	0.88	0.86	0.87
Random Forest	0.86	0.75	0.75	0.86	0.8	0.88	0.83
K nearest	0.5	0.86	0.75	0.67	0.6	0.75	0.69

6.CONCLUSION

With good precision, recall, F1 score, and accuracy for both Class 0 and Class 1, Logistic Regression is the most balanced and successful model overall, according to the performance measures. Random Forest has good performance but trails behind Logistic Regression, particularly in Class 1, with intermediate precision and recall. Decision Tree performs well in recall for Class 1 but poorly for Class 0, resulting in a lower F1 score and accuracy compared to Logistic Regression. When it comes to Class 0, in particular, K-Nearest Neighbor (KNN) performs the worst, exhibiting lesser precision and recall, which results in the lowest accuracy. Consequently, the best model for this classification problem is logistic regression, while decision trees and random forests might also work depending on goals.

7.ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without mentioning about the people who made it possible, whose constant guidance and encouragement has crowned the efforts with success. I would like to thank almighty God who gave knowledge and strength to complete this project successfully.

With deep gratitude, I profoundly thank our CEO & Secretary **Dr P. KRISHNA KUMAR**, **MBA, PhD, D.Litt.** for permitting me to undertake this work in Nehru College of Management.

I express my deep sense of gratitude to Principal, **Dr.R MOSES DANIEL, MBA, PhD, PGDCA, PGDBI,** and **Dr.M.SENGALIAPPAN**, **MCA, MPhil, ME, Ph.D.,** Head of the Department, Department of Computer Applications for providing me with all the necessary facilities, which helped me, a lot in the successful completion of this project.

It is a great pleasure for me to acknowledge the assistance of my project guide **Dr.M.SENGALIAPPAN,** Associate professor ,Head of the Department ,Department of Computer Applications, Nehru College of Management, Coimbatore, for his valuable advice and kind support during the period of this project.

8.REFERENCES:

[1] "*ML Methods for Crop Yield Prediction and Estimation: An Exploration,*" International Journal of Engineering and Advanced Technology, vol. 9 no. 3, 2020, M. Alagurajan and C. Vijayakumaran

[2] "*Prediction of Crop Yield Using SVM Approach with the Facility of E-MART System,*" Easychair 2020, P. Kumari, S. Rathore, A. Kalamkar, and T. Kambale.

[3] "*A microcontroller based machine vision approach for tomato grading and sorting using SVM classifier,*" Microprocessors and Microsystems, vol. 76, pp. 103090, 2020, S. D. Kumar, S. Esakkirajan, S. Bama, and B. Keerthiveena

[4] "*Crop yield prediction by modified convolutional neural network and geographical indexes,*" by P. Tiwari and P. Shukla, International Journal of Computer Sciences and Engineering, vol. 6, no. 8, pp. 503-513, 2018.

[5] "*Crop Yield Prediction Analysis using Feed Forward and Recurrent Neural Network,*" International Journal of Innovative Science and Research Technology, vol. 5, no. 5, pp. 1092-1096, 2020, author P. Sivanandhini and author J. Prakash.

[6] Nandhini and J. G. Shankar, "*Use of machine learning based on seed to predict crop growth,*" Ictact journal on soft computing, vol. 11, no. 01, 2020

[7] "*Crop prediction based on geographical and climatic data using machine learning and deep learning,*" doctoral dissertation, BRAC University, 2018; A. A. Alif, I. F. Shukanya, and T. N. Afee.

[8] "*A robust deeplearning-based detector for real-time tomato plant diseases and pests' recognition,*" Sensors, vol. 17, no. 9, pp. 2022, 2017, by A. Fuentes, S. Yoon, S.C.Kim,andD.S.Park.

[9] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "*Deep CNN-LSTM model for county-level soybean yield prediction,*" Sensors, vol. 19, no. 20, pp. 4363, 2019.
[10] "Hybrid prediction strategy to predict agricultural information," Applied Soft Computing, vol. 98, pp. 106811, 2021, by K. A. Shastry and H. A. Sanjay.

[10]"*Crop Yield and Fertilizer Recommendation Prediction Using Machine Learning Algorithms,*" International Journal of Engineering Applied Sciences and Technology, vol. 4, no. 5, pp. 371-376, 2019, D. A. Bondre and S. Mahagaonkar.