Estimating Crop Yields Using Machine Learning Techniques

T. Anuradha[1], B. Sai Naga Himaja[2], M. Sai Manish Varma[3], N. Sandeep[4], *Dr. P M Manohar[5]

anuradha.tutika@raghuenggcollege.in, 19981a05e4@raghuenggcollege.in ,20985A0524@raghuenggcolegeg.in, 20985A0526@raghuenggcollege.in, manohar.pattisapu@raghuenggcollege.in

[1] Assistant professor, Department of Computer Science Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh.

[2,3,4] Fourth year BTech Students, Department of Computer Science Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh.

[*5] Associate Professor, Department of Computer Science Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh.

Abstract

India's economy heavily relies on agriculture, with a significant portion of the population either directly involved in farming or trading the end products. Agricultural production of food grains in India is vast and has a tremendous impact on the nation's economy. Farmers are eager to know the yield they can expect, and this study attempts to quantify the technological factors that affect yields. By analysing factors such as season, area, temperature, humidity, moisture, soil type, crop type, and nutrient levels, the study aims to predict future crop yields accurately. Additionally, the model uses data on crop production in different districts and years to train and test various machine learning (ML) Techniques like Decision tree, Linear regression, K-Nearest Neighbors along with a deep learning architecture. The above listed models are compared based on different performance metrics such as RMSE and MAE etc. This study's results can assist agronomists a way to increase their incomes along with country economy by crop yield prediction with a fertilizer need to be used for that given area.

Keywords--- Agriculture, crop-prediction, Regressions.

I. INTRODUCTION

Agriculture is a crucial sector in the country's economy and sustainability. With the increase in population, the importance and demand for agriculture are rising. Therefore, to maintain a sustainable balance we need to improve the growth of agriculture production. The crop yield is the primary measure of agricultural outcomes, which mainly depends on various agricultural factors. The agricultural process is broadly categorized into two stages: pre-sowing and post-sowing. In the pre-sowing stage, farmers must analyse the climate for the sowing period, assess soil nutrients, determine soil texture, and select an appropriate crop based on the soil type. The post-sowing stage is more critical and involves pest management, irrigation, fertilizer management, weed management, usage of tools, and proper timing of harvest. Each step in the pre- and post-sowing process has a defined role in determining crop yield, and following all the steps carefully can lead to higher yields. Machine learning regression techniques can be employed to predict crop yields by thoroughly analysing previous data. This approach enables us to suggest better crops to farmers that can yield higher profits. The challenge, however, is to build an efficient model that can accurately predict the crop yield. To address this, various algorithms can be tried and tested to determine which one has the lowest error rate and loss. Once the best model is identified, the yield of the specific crop can be predicted. This paper compares two or more models and predicts the output using the most effective model.

The paper is organized into several sections. The first section provides an introduction to the use of machine learning techniques for estimating crop yields. The second section is a literature review on the topic. The third section describes the methodology used in the study, including the dataset explanation, data pre-processing, proposed model architecture, and the different techniques used such as KNN, decision tree regression, linear regression, and LSTM architecture. The fourth section presents the results and their analysis. The last section is the conclusion, which also discusses the future scope of the study.

II. LITERATURE SURVEY

Shruti Mishra et al. (2018) [3] employed multiple data mining classifiers including Instance Based Learner (IBK), Locally Weighted Learning (LWL), LAD Tree,

and J48 to forecast crop production. They analysed various factors such as district name, crop year, season name, crop name, and area. These factors were pre-processed and classified using the WEKA explorer. The prediction error rate was calculated using Root Mean Squared Error and Mean Absolute Error as shown in table 1.

Algorithm	RMSE	MAE
J48	**0.2773**	**0.1101**
LWL	0.3209	0.2213
LAD Tree	0.4127	0.1997
IBK	0.104	0.104

Table. 1 Error Rate

Kumar. K et al [8], for addressing the problems in selecting crop he introduced a model called a Novel ML. For identifying the suitable crop along with increase in its result and net rate of crop yield by using a method called CSM (Crop Selection Method).

M. M. Rahman et al developed ML (machine learning) model' for predicting the production rate of rice in Bangladesh. It is a place where the soil condition is not uniform. Here the models were trained initially based on the parameters like the crop yield rate of the previous climatic conditions and environments. These are the factors they considered to predict the accuracy.

G. Deepti and udayan together developed forecasting models for rainfall based on their classification and NBS (Naive Bayes Approach), KNN (k-nearest neighbour), RTA (Regression Tree Algorithm) & 5-10-1 Pattern recognition NN (Neural Network). The obtained accuracy based on their results to corresponding techniques are respectively 78.89%, 80.7%, 80.3%, 82.1%.

S. Veenadhari, Bharat Mishra, CD singh in 2011 [6] said that it is possible to measure crop productivity of soyabean accurately by using (decision tree) algorithm. Soyabean has been the chief crop of MP (Madhya Pradesh) from the last two decades there is no good yield productivity even the cultivation land is increased. Climatic factors like relative humidity plays major role in soyabean crop productivity, so with the help of (decision tree algorithm) we can consider all these parameters along with previous year's soya crop production. Then it is accurately possible to predict the yield of soyabean. After using decision tree algorithm, we can see that there is a clear correlation between both soyabean crop production and the climatic factors. This algorithm can only be used for this crop and cannot be used for any other crops in different places.

III. METHODOLOGY

The ability to estimate crop yield in advance is highly beneficial for farmers and the agriculture industry, allowing for improved planning and decision-making in crop management. This can be achieved using powerful algorithms that predict both crop production and the type of fertilizer required. The proposed methodology explores various regression techniques, including decision tree, linear regression, and K-Nearest Neighbors, while also comparing these machine learning models to a deep learning architecture known as LSTM.

A. Data Set Explanation

This model employs two datasets [2] from kaggle. The first dataset named as “cropyield” as shown in Figure 1, includes seven factors, such as state name, district name, crop year, season, crop, area, and production, which is used to train different machine learning techniques for predicting the crop yield and have 246091 records in it. The second dataset named as “fertilizer prediction”, represented in Figure 2, is used to recommend the fertilize name for certain crop that should be used for obtaining a better yield of the crop. Fertilizer prediction is based on nine factors, including temperature, humidity, moisture, soil type, crop type, nitrogen, potassium, phosphorous, and Fertilizer name.

State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254	2000
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2	1
Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102	321
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176	641
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720	165
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Coconut	18168	65100000
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Dry ginger	36	100
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sugarcane	1	2
Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sweet potato	5	15

Figure. 1 Sample crop data

Temparature	Humidity	Moisture	Soil_Type	Crop_Type	Nitrogen	Potassium	Phosphorous	Fertilizer_Name
26	52	38	Sandy	Maize	37	0	0	Urea
29	52	45	Loamy	Sugarcane	12	0	36	DAP
34	65	62	Black	Cotton	7	9	30	14-35-14
32	62	34	Red	Tobacco	22	0	20	28-28
28	54	46	Clayey	Paddy	35	0	0	Urea
26	52	35	Sandy	Barley	12	10	13	17-17-17
25	50	64	Red	Cotton	9	0	10	20-20
33	64	50	Loamy	Wheat	41	0	0	Urea
30	60	42	Sandy	Millets	21	0	18	28-28

Figure. 2 Sample Fertilizer data

B. Data Pre-processing

Data processing comprises a series of steps for verifying, arranging, converting, amalgamating, and extracting data to generate an appropriate output. It is imperative to thoroughly document the processing methods to ensure data's veracity and usefulness. Since our dataset involves text data in a few columns, the foremost task is to convert this textual data into numerical format.

To divide the dataset into training and testing subsets, the 'train_test_split' function from the 'sklearn' library can be employed. This function allows us to partition the data based on a specified split ratio [2]. It is advisable to assign a split ratio of 80:20, which corresponds to 80% of the data being used for training and 20% for testing.

C. Proposed Model Architecture

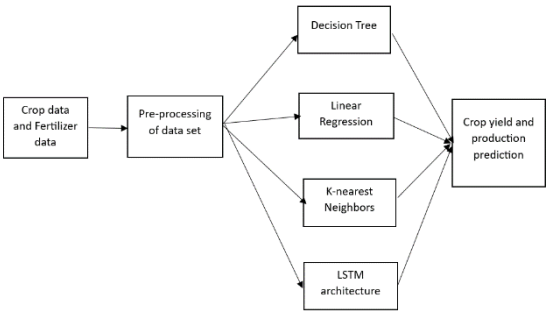

Figure. 3 Architecture model for proposed system

D. Implementation Techniques

a) Linear Regression

Linear Regression is a type of machine learning algorithm that belongs to the supervised learning category and is commonly used to solve regression problems [10]. It predicts an output value based on predicator variables and models the correlation between these predictors and the expected output. Regression models vary depending on the relationship between predictors and the predicted output, and the number of predictor variables used. The most basic mathematical relationship between two variables is a linear relationship, which is often used in cause-and-effect scenarios where the predictor variable represents the cause and the predicted output represents the effect.

One drawback of using Linear Regression is that it may not be suitable for handling complex datasets that exhibit non-linear relationships between their features. It is also revealed that linear regression did not produce satisfactory results for any of the datasets we examined, likely due to the non-linear nature of the data.

b) Decision Tree Regression

The Decision Tree algorithm constructs a tree-like structure for both classification and regression models [10]. This involves splitting the dataset into smaller subsets, while simultaneously generating a related decision tree. As a supervised learning algorithm, it caters to both categorical and continuous output variables. The nodes in the tree can either be conditions or results, known as decision nodes and end nodes, respectively.

Decision tree regression analyses an object's characteristics and creates a tree-based model to predict continuous outcomes for future production. When we talk about continuous output, we are referring to results that are not characterized by discrete or distinct values or numbers. Instead, they can take on a range of values within a certain range or interval.

One of the limitations is that it may overfit the data and create overly complex models that do not generalize well to new data. Another limitation is that it may not handle well datasets with missing values or outliers. Finally, decision tree models are not suitable for datasets with high dimensionality, as the number of possible splits and combinations of features can become too large to handle efficiently.

c) K-Nearest Neighbors

The K-Nearest Neighbors (KNN) method is a widely used data mining technique for crop yield estimation [5]. In this method, each feature related to the crop, such as soil type, weather conditions, and fertilizers used, is treated as a distinct dimension in a space. The crop instances are represented as points in this space, with the value of each feature corresponding to the coordinate of the point in that dimension.

To predict the yield of a new crop instance, the KNN algorithm calculates the distance between the new instance and all the instances in the training dataset using a suitable distance metric, such as Euclidean distance [5]. The algorithm then selects the k closest instances based on the smallest distances, and determines the most common yield among them. This yield is then assigned to the new instance as the predicted yield.

This method can be used with textual datasets as well by converting the textual data into numerical values through techniques like word embeddings or one-hot encoding. The KNN algorithm can then be applied to predict the yield of a new crop instance based on the textual features.

d) LSTM (Long Short-Term Memory)

In crop yield prediction, the Long Short-Term Memory (LSTM) architecture which belongs to deep learning domain is a popular choice for processing textual data as input. In the context of crop yield prediction, an LSTM model can be trained on textual data such as weather reports, soil conditions, and other relevant information to predict crop yield. The model is trained to learn patterns in the input data and to use these patterns to make accurate predictions about future crop yields.

The RMSE value of 0.018 indicates that the LSTM model has performed well in predicting crop yield. The small difference between the predicted and actual crop yields suggests that the LSTM model is effective in accurately predicting crop yields using textual input data. This approach shows potential for improving decision-making related to crop management,

particularly for farmers. Nevertheless, the use of LSTM for crop yield prediction using textual input data appears to be a promising approach that could lead to significant advancements in agriculture.

IV. RESULTS AND ANALYSIS

By using textual data sets precisely compared to the existed models we estimated the yield of crop production, in this paper. We considered RMSE, MAE and R2 as metrics for evaluating the performance. The below table explains about each regression model's error rate in three various performance measures.

Machine learning Models	RMSE (Root Mean Square Error)	MAE (Mean Absolute Error)	R2 (R-Squared Error)
Decision Tree Regression	0.0671	0.0074	-1.9116
Linear Regression	0.0716	0.0078	-4.5278
K-Nearest Neighbors	0.0570	0.0066	0.1821

Table. 2 Error Rate Comparison

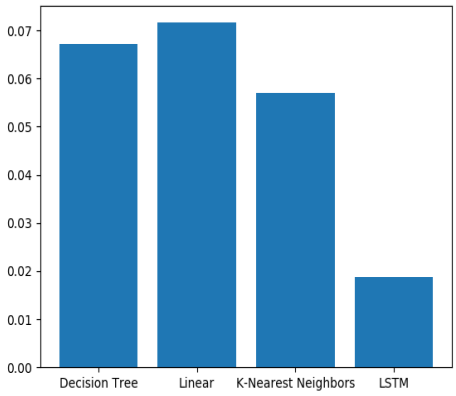

Figure. 4 RMSE (Root Mean Square Error) Comparison Graph

Table 2 shows that the Linear Regression model has a RMSE value of 0.0716 and a MAE value of 0.0078. The Decision Tree model has an RMSE value of 0.0671 and an MAE value of 0.0074. The KNN model outperforms both models with the lowest RMSE value of 0.0570, the lowest MAE value of 0.0066 and R2 error value is 0.1821. These results indicate that the KNN model is the most accurate and reliable model for estimating crop yields and production. Also, LSTM architecture model has an RMSE value of 0.01873.

Figure 4 illustrates a comparison graph that evaluates the root mean square error for four different models, including Decision Tree Regression, Linear Regression, K-Nearest Neighbors, and LSTM. The graph allows for a comparison of the performance of these models in predicting crop yield using textual input data.

Here, in order to predict crop yields, the KNN model can be preferred over linear regression and decision tree. The accuracy of the KNN model can help farmers and agricultural scientists in making informed decisions about managing crop and its strategical production. We used LSTM after comparing KNN model with DL (deep learning) architecture. By analysing both the results KNN is outperformed by LSTM. Overall, it increases the production and efficiency of agricultural practices when we use ML models, which leads to better crop yields and increased food production. At the same time, deep learning architecture increases the accuracy.

V. CONCLUSION AND FUTURE SCOPE

In conclusion, the project of predicting crop yields using machine learning models is highly promising for farmers and agricultural scientists. The three models, Linear Regression, Decision Tree, and KNN, can provide accurate predictions for crop yields and production. Among the three models, the KNN model has the lowest RMSE and MAE values, indicating that it is the most accurate and reliable model for predicting crop yields. The LSTM architecture model also increased the accuracy and have less error rate while we performed separately, it indicated to be the best comparing with those three machine learning models. The use of machine learning models can help farmers and agricultural scientists make informed decisions about crop management and production strategies, leading to better crop yields and increased food production.

In terms of future scope, the project can be extended by incorporating additional features such as soil and weather data, which can further improve the accuracy of the models. The project can also be integrated with mobile applications and sensors, allowing farmers to access real-time crop yield predictions and make timely decisions. Additionally, the project can be expanded to cover a wider range of crops and regions, enabling farmers and agricultural scientists to make crop yield predictions for various types of crops in different regions. Overall, the project has significant potential to contribute to the development of precision agriculture and improve the efficiency and productivity of agricultural practices.

REFERENCES

[1] N. K. Choudhary, S. Sree Laya Chukkapalli, S. Mittal, M. Gupta, M. Abdelsalam and A. Joshi, "YieldPredict: A Crop Yield Prediction Framework for Smart Farms," 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 2340-2349, doi: 10.1109/BigData50022.2020.9377832.

[2] T. Nikhil Yadav, Dr. R. Obulakonda Reddy, Ram Chandra Prasad, Pradeep Gopal (2020)). Crop Yield and Fertilizers Prediction Using Decision Tree

Algorithm, International Journal of Engineering Applied Sciences and Technology (IJEAST), ISSN: 2455-2143.

[3] Mishra, S., Yadav, S. S., Tiwari, R., & Singh, U. (2018). Forecasting of crop production using data mining techniques: A case study of Madhya Pradesh, India. Journal of King Saud University-Computer and Information Sciences, 30(1), 57-70.

[4] Dharani, M & Thamilselvan, R & Natesan, P & Kalaivaani, PCD & Santhoshkumar, S. (2021). Review on Crop Prediction Using Deep Learning Techniques. Journal of Physics: Conference Series. 1767. 012026. 10.1088/1742-6596/1767/1/012026.

[5] B. Lakshmi Praveena, A. Bhumika, A. Sindhuja, & B. Tejaswini Mallika (2022). Crop Prediction Using KNN Algorithm, International Journal of Advance Research and Innovative Ideas in Education (IJARIIE), ISEES(O)-2395-4396.

[6] Veenadhari, S., Mishra, B., & Singh, C. D. (2011). Decision tree algorithm for prediction of soybean productivity in relation to climate parameters. International Journal of Earth Sciences and Engineering, 4(6), 1087-1091.

[7] Alexandros Oikonomidis, Cagatay Catal & Ayalew Kassahun (2023) Deep learning for crop yield prediction: a systematic literature review, New Zealand Journal of Crop and Horticultural Science, 51:1, 1:26, DOI: 10.1080/01140671.2022.2032213.

[8] Kumar, K., Kumar, R. R., Kumar, A., & Soni, R. (2019). Crop selection based on multiple attributes using novel machine learning model. Journal of Ambient Intelligence and Humanized Computing, 10(3), 1019-1029.

[9] N. Mounika, Dr.D.Vivekananda Reddy (2021). Crops Yield Prediction and Efficient Use of Fertilizers Using Machine Learning, Journal of Emerging Technologies, and Innovative Research (JETIR), ISSN-2349-5162.

[10] D.Bhanu Kiran, J.Priyanka, K.Sri Poojitha, & Afridi Khan (2020). Crop Yield Prediction Using Regression, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056.