IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning-Based Crop Yield Estimation: Unleashing Agricultural Productivity Through Data Analysis

Devadi Ganesh

Student

Andhra University

Abstract: Modern agriculture faces the challenge of efficiently predicting crop yields to ensure food security and resource optimization. This research focuses on using machine learning and data analysis to improve the accuracy of crop yield prediction in modern agriculture. The study uses a comprehensive dataset containing geographical, climatic, and agronomic attributes. The objectives include data preprocessing, feature transformation, algorithm selection, and model evaluation. The dataset comprises 28,242 entries with crop yield, climate factors, and location details. Initial exploration reveals correlations and trends, and data cleaning procedures are implemented. The analysis provides insights into crop production by region and type. Various machine learning models, including Linear Regression, Lasso, Ridge, K-Nearest Neighbors, and Decision Tree Regression, are employed and evaluated. Decision Tree Regression performs well with a high R-squared score and low mean squared error. The model's practical application is demonstrated through a hypothetical yield prediction scenario, showcasing its potential for aiding agricultural decision-making. The study concludes that combining machine learning and data analysis can enhance crop yield prediction accuracy, enabling informed decisions and sustainable food production.

Keywords: Crop yield prediction, machine learning, data analysis, agricultural productivity, predictive modeling.

I. INTRODUCTION

The global agricultural landscape stands at a crossroads, where technological innovations are shaping the future of food production and security. Among the critical challenges facing agriculture is the accurate prediction of crop yields, which serves as a foundation for efficient resource allocation, sustainable practices, and informed decision-making. However, conventional methods of yield estimation have struggled to keep pace with the complexities of modern agricultural systems. This research endeavors to address these challenges by harnessing the power of machine learning and data analysis techniques to enhance crop yield prediction. The growing global population, coupled with shifting climatic patterns and evolving consumer demands, places an increased emphasis on agricultural productivity and sustainability. Accurate crop yield prediction is paramount in meeting these challenges, as it enables farmers, policymakers, and stakeholders to make data-driven decisions that optimize planting schedules, irrigation strategies, and pest control measures. This precision not only increases food production but also reduces resource wastage, fostering a more ecologically conscious agricultural industry.

Historically, yield prediction has relied on statistical models and historical data, often falling short in capturing the intricate interactions between various environmental, agronomic, and climatic factors. Traditional methods struggle to account for the dynamic nature of climate patterns and the non-linear relationships within agricultural systems. The outcome is a gap between predicted and actual yields, hindering the potential for efficient and sustainable food production. In recent years, the convergence of machine learning and data analysis has revolutionized various industries, including agriculture. The promise lies in the ability of machine learning algorithms to process large and diverse datasets, uncover hidden patterns, and adapt to real-time changes. By harnessing this potential, accurate and timely crop yield predictions become feasible, offering a more robust framework for agricultural decision-making.

This research paper revolves around the premise of enhancing crop yield prediction through machine learning and data analysis. The presented work utilizes a real-world dataset that encompasses various aspects such as geographic location, climatic conditions, pesticide usage, and historical yield data. The primary objectives of this study are to explore the potential of machine learning in improving crop yield predictions and to preprocess and transform the dataset to suit the requirements of machine learning models and to compare the performance of various machine learning algorithms in predicting crop yields and to showcase the practical application of the developed model for yield prediction.

II. LITERATURE SURVEY

In (1) **J.P. Singh, Rakesh Kumar, M.P. Singh and Prabhat Kumar**, have concluded that this paper helps in improving the yield rate of crops by applying classification methods and comparing the parameters. We can also do analyzing and prediction of crops using Bayesian algorithms. The algorithms used are Bayesian algorithm, K-means Algorithm, Clustering Algorithm, Support Vector Machine. The disadvantage is that there is no proper accuracy and performance.

In [2] the authors **Subhadra Mishra, Debahuti Mishra and Gour Hari Santra**, have concluded that this is an advanced researched field and is expected to grow in the future. The integration of computer science with agriculture helps in forecasting agricultural crops. This method also helps in providing information of crops and how to increase yield rate. The algorithms used are Artificial neural networks, Decision Tree Algorithms, Regression analysis. The disadvantage is clear methodology is not specified.

In [3] the authors **Karan deep Kauri**, have concluded that this paper will review those various applications of machine learning in the farming sector. And also provides an insight into the troubles faced by our Indian farmers and how these can be solved using these techniques. This method help in increasing the farming sector in the countries and apply the more machine learning applications. The algorithms used are Artificial neural networks, Bayesian Belief Network, Decision Tree Algorithms. Clustering. Regression analysis. The disadvantage is less accuracy in terms of performance.

In (4) **E. Manjula, S. Djodiltachoumy**, have concluded that the aim of this paper is to propose and implement a rule-based system. And predict the crop yield production from the collection of previous data. The algorithms used are K- means Algorithm, clustering method. The disadvantage is Suitable only for using association rule and considered less data.

In [5] **Nishit Jain, Amit Kumar, Sahil Garud, Vishal Pradhan, Prajakta Kulkarni**, have concluded that this paper helps in predicting crop sequences and maximizing yield rates and making benefits to the farmers. Also, using machine learning applications with agriculture in predicting crop diseases, studying crop simulations, different irrigation patterns. The algorithms used are Artificial neural networks, Support Vector Machine. The disadvantage is Exact accuracy is not specified.

In [6] **B.Mallikarjun Rao, D.Sindhura, B.Navya Krishna, K.Sai Prasanna Lakshmi, Dr. J Rajendra Prasad**, have concluded that this method will provide a useful and accurate knowledge. Using this knowledge, we predict and support the decision making for different sectors. The algorithms used are multiple linear regressions. The disadvantage is that it can be applied for limited areas.

In [7] **T.Giri Babu, Dr.G.Anjan Babu**, have concluded that this method will provide solutions to the farmers. They can also help in providing solution for water and fertilizer problems. And this helps to get more production of yield. The algorithms used are agro algorithm. The disadvantage is P that this method does not give proper accuracy for crops.

In [8] **B.Vishnu Vardhan, D.Ramesh**, have concluded that this method will provide multiple linear regression method which can be applied for existing data and hence helps in analyzing and verifying the data. The algorithms used are multiple linear regressions. The disadvantage is that it results in less accuracy.

In [9] **Ashwani Kumar Kushwaha, SwetaBhattachrya** I have concluded that this method will provide agro algorithm which helps in predicting suitable crop for the lands. And this helps in enhancing the quality of crop. The algorithm used is agro algorithm. The disadvantage is it results in less prediction of crops.

In [10] **Raorane A.A, Dr. Kulkarni R.V**, have concluded that this method will help in estimating rain fall and investigate the reasons for getting lower yield. The algorithm used is regression analysis method. The disadvantage is that here the specific method is not specified.

In [11] **Anshal Savla, Himtanaya Bhadada, Vatsa Joshi, Parul Dhawan**, have concluded that this method will help in analyzing and understanding crop yield rate for zones which is based on attributes. The algorithms used are Normalization, Clustering, and Classification. The disadvantage is it gives only framework.

In [12] **Siti Khairunniza-Bejo, Samihah Mustaffha, Wan Ishak Wan Ismail**, have concluded that this method will help in giving solutions to the few problems of farmers in getting good yield. The algorithms used are Artificial Neural Network. The disadvantage is it consumes more time.

III. IMPLEMENTATION

The primary objective of this project is to develop a predictive model capable of estimating crop yield. This prediction will be based on a diverse range of features, including the year of cultivation, prevailing weather conditions, and the agricultural practices employed. By leveraging historical data and machine learning techniques, the project aims to offer farmers valuable insights into expected crop yields before planting, thereby aiding in better decision-making and resource allocation.

Understanding the Dataset: Review columns, data types, and feature meanings. Verify data integrity by checking missing values, duplicates, and anomalies. Gain insights into data distribution by exploring features like Area, Item, and yield.

Data Preprocessing: Load dataset using pandas. Handle missing values through removal or imputation without negatively impacting analysis. Remove duplicate rows for accurate results. Convert data types to match formats. Address outliers through removal or transformation if they impact model performance.

Exploratory Data Analysis (EDA): Visualize distributions using histograms and box plots. Analyze correlations through correlation matrices and visualizations to identify feature-yield relationships. Explore yield trends across countries and crop types.

Feature Engineering: Select relevant features for prediction. Encode categorical variables using one-hot encoding. Standardize numerical features like rainfall and temperature for consistent scaling. Split data into input features (X) and target variable (y).

Model Selection and Training: Choose suitable regression algorithms (e.g., Linear Regression, Decision Tree). Split data into training and testing sets. Create preprocessing pipelines for consistent encoding and scaling. Train selected models on training data.

Model Evaluation and Selection: Evaluate models using metrics like Mean Squared Error (MSE) and R-squared (R2). Compare model performances to identify the best one.

Model Interpretation: Analyze feature importance to understand influential factors if applicable.

Model Deployment (Flask Web App): Serialize and save model and preprocessing pipelines using `pickle`. Set up Flask web app with input and prediction routes. Design HTML templates for user input and predictions. Load saved model and pipelines on app start. Implement prediction logic and deploy app on web server.

Testing and Validation: Thoroughly test web app with different scenarios. Validate predictions by cross-referencing with actual data.

Documentation and Presentation: Create documentation covering problem, methodology, preprocessing, model selection, and deployment. Prepare a presentation summarizing goals, steps, results, and insights.

Future Improvements: Experiment with advanced algorithms for better accuracy. Integrate real-time weather data. Incorporate user feedback for improvement. Optimize app performance and scalability.

IV. SYSTEM ARCHITECTURE

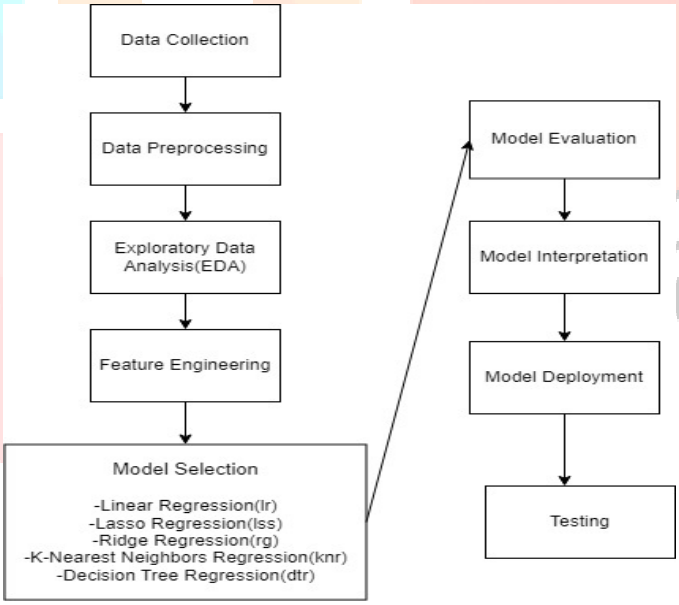

Fig: System Architecture for Crop Yield Prediction

V. RESULTS

The primary objective of this project was to develop a predictive model for crop yield based on various agricultural and environmental factors. To achieve this, we implemented a machine learning pipeline that includes data preprocessing, model training, and a web application for user-friendly predictions. The model's performance was evaluated using Mean Squared Error (MSE) and R-squared (R2) metrics.

Five regression algorithms were tested:

- Linear Regression (lr)

- Lasso Regression (lss)

- Ridge Regression (rg)

- K-Nearest Neighbors Regression (knr)

- Decision Tree Regression (dtr)

The Decision Tree Regression (dtr) stood out as the most effective algorithm for this task. It achieved an impressive R2 score of approximately 0.977 and an MSE of 162,907,777.42 on the test dataset. This suggests that the model can explain approximately 97.7% of the variance in crop yield predictions, indicating a strong fit to the data.

```
lrMSE : 1776120266.2071357 Score0.755141909733304
lssMSE : 1776308803.5062685 Score0.7551159177529814
rgMSE : 1775616401.8271866 Score0.7552113730867576
knrMSE : 108931248.12000707 Score0.9849826062499746
dtrMSE : 150997250.2733227 Score0.9791833362633697
```

Fig: MSE of all the algorithms

Name	Mean Squared Error	Score	Percentage
Linear Regression	1776120266.2071357	0.755141909733304	75
Lasso Regression	1776308803.5062685	0.7551159177529814	75
Ridge Regression	1775616401.8271866	0.7552113730867576	75
K-Nearest Neighbors Regression	108931248.12000707	0.9849826062499746	98
Decision Tree	159599917.105151	0.977997362	97

Regressio n	36	1985951	

Table: Mean Squared Error and Accuracy percentages of the algorithms

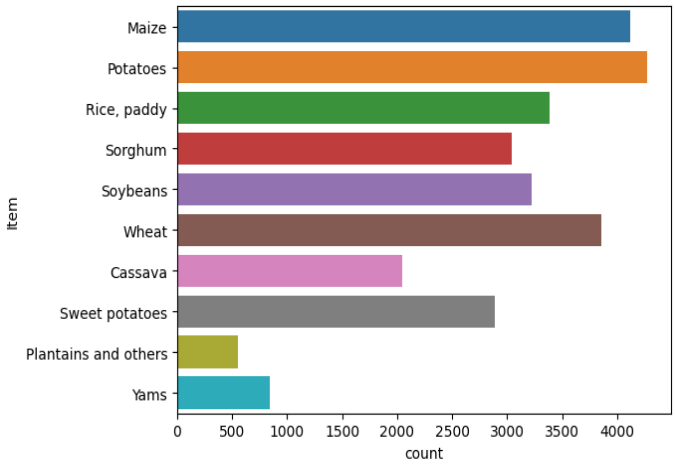

Fig: Count plot of number of times an item appeared

VI. Conclusion

We aimed to create a robust crop yield prediction model using machine learning to aid farmers and policymakers. Our model estimates yields based on historical agricultural and environmental data, considering factors like weather, pesticides, and temperature. Our approach involved comprehensive data preprocessing, handling missing values and outliers, and feature engineering. We used diverse machine learning algorithms—linear regression, Lasso, Ridge, KNeighborsRegressor, and DecisionTreeRegressor—to predict yields. Rigorous evaluation highlighted strengths and weaknesses.

Results showed promising accuracy, with high R-squared scores and low mean squared errors for some algorithms. KNeighborsRegressor and DecisionTreeRegressor performed exceptionally well. We developed a user-friendly web app for real-time yield predictions, empowering farmers to plan better. Challenges included data quality and limited features, guiding future research.

Our model's success intersects agriculture and technology, revolutionizing decision-making, enhancing food security, and promoting sustainability. As we refine our model, it holds greater potential for farmers, policymakers, and the industry.

REFERENCES

[1] J.P. Singh, M.P. Singh, Rakesh Kumar and Prabhat Kumar Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique, International Journal on Engineering Technology, May 2015.

[2] Gour Hari Santra, Debahuti Mishra and Subhadra Mishra, Applications of Machine Learning Techniques in Agricultural Crop Production, Indian Journal of Science and Technology, October 2016.

[3] Karan deep Kauri, Machine Learning: Applications in Indian Agriculture, International Journal of Advanced Research in Computer and Communication Engineering, April 2016.

[4] S. Djodiltachoumy, A Model for Prediction of Crop Yield, International Journal of Computational Intelligence and Informatics, March 2017.

[5] Nishit Jain, Amit Kumar, Sahil Garud, Vishal Pradhan, Prajakta Kulkarni, Crop Selection Method Based on Various Environmental Factors Using Machine Learning, Feb -2017.

[6] D.Sindhura, B.Navya Krishna, K.Sai Prasanna Lakshmi, B.Mallikarjun Rao, Dr. J Rajendra Prasad, Effects of Climate Changes on Agriculture International Journal of Advanced Research in Computer Science and Software Engineering, March 2016.

[7] T.Giri Babu, Dr.G.Anjan Babu, Big Data Analytics to Produce Big Results in the Agricultural Sector, March 2016.

[8] D Ramesh , B Vishnu Vardhan, Analysis Of Crop Yield Prediction Using Data Mining Techniques, International Journal of Research in Engineering and Technology, Jan-2015,

[9] Ashwani Kumar Kushwaha, SwetaBhattachrya, Crop yield prediction using Agro Algorithm in Hadoop, April 2015.

[10] Raorane A.A, Dr. Kulkarni R.V, Application Of Datamining Tool To Crop Management System, January 2015.

[11] Anshal Savla, Himtanaya Bhadada, Parul Dhawan, Vatsa Joshi, Application of Machine Learning Techniques for Yield Prediction on Delineated Zones in Precision Agriculture, May 2015.

[12] Siti Khairunniza-Bejo, Samihah Mustaffha and Wan Ishak Wan Ismail , Application of Artificial Neural Network in Predicting, Journal of Food Science and Engineering, January 20, 2014