IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13903

Agriculture Crop Yield Prediction

SHRIRAKSHA I P[1], PROF. SANDEEP N K[2]

Department of Masters of Computer Application, Vidya Vikas Institute of Engineering & Technology Mysore, Karnataka, India.[1]

Assistant Professor, Department Of Masters Of Computer Application, Vidya Vikas Institute Of Engineering & Technology, Mysore, Karnataka, India [2]

Abstract: In the quest to enhance agricultural productivity, predicting crop yield plays important part in optimizing resource allocation and planning. This study explores the uses of machine learning model to forecast crop yield, leveraging various models to analyse and interpret historical data. By integrating parameter like crop type, temperature, rainfall, and pesticide use, Machine learning techniques yield precise outcomes. predictions that support decision-making processes in agriculture. The results demonstrate the potential of these advanced analytical methods to provide actionable insights, improve yield forecasting accuracy, and ultimately contribute to sustainable agricultural practices.

I. INTRODUCTION

Agriculture remains a crucial sector for society, since it is the primary source of food production. Despite these advancements, numerous countries still struggle with hunger due to food shortages and the difficulties brought about by a rapidly increasing population. Increasing food production is essential to eradicating famine and ensuring food security for all. Achieving food security and reducing hunger by 2030 are key goals set by the United Nations. Consequently, crop protection, land assessment, and crop yield prediction have become increasingly significant in global food production efforts.

Machine learning, with its diverse strategies and methodologies, it is crucial in these regions. By employing a range of machine learning methods, data scientists can uncover meaningful patterns and insights from agricultural datasets.

In this context, the agent, representing the machine, learns to achieve a specific goal within a complex and uncertain environment, with the environment providing rewards based on the agent's actions.

Problem statement:
Crop yield prediction is a critical challenge in agriculture, significantly impacted by variables such as weather conditions (rainfall, temperature, etc.) and pesticide use. Having precise historical data on crop yields is vital for ensuring knowledgeable decision-making in agricultural risk management and forecasting future yields Given the complexity of predicting crop yields, machine learning techniques are applied to address these challenges effectively.

II. LITERATURE SURVEY

Agriculture is a pivotal sector for India's economy, contributing 18% to the Gross National Product (GNP) and employing 50% of the population. Despite longstanding agricultural practices, crop yields often fall short due to various influencing factors. Meeting the needs of approximately 1.2 billion people requires optimizing crop yields. Factors such as soil type, precipitation, seed quality, and technical limitations significantly impact crop productivity. To address these challenges, adopting advanced technologies is crucial, moving beyond traditional methods.

This paper focuses on using data mining methods for crop yield prediction, analysing an agricultural dataset. Various classifiers, including J48, LWL, LAD Tree, and IBK, are assessed using the WEKA tool. Performance is measured by and evaluated using Root Mean Squared Error (RMSE), mean absolute error (MAE), and Relative Absolute Error (RAE). Lower error values indicate higher accuracy. The study compares these classifiers based on their performance.

In India, food production is largely dependent on cereal crops such as rice, wheat and pulses. The output of rice-growing regions is heavily influenced by climatic conditions, with droughts potentially reducing yields. Improving techniques to predict performance of crop yield under different climatic conditions can aid farmers and stakeholders in making better agronomic and crop decisions. Machine learning methods present significant potential solutions for enhancing yield predictions across different climatic conditions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13903

This paper reviews the use of machine learning methods for rice cropping areas in India. It discusses experimental results using the SMO classifier on data from 27 districts in Maharashtra, sourced from Indian Government records. Key parameters include precipitation, temperature variations, and crop evapotranspiration for the Kharif season from 1998 to 2002. Performance metrics such as MAE, RMSE, RAE, and RRSE were employed to assess results, revealing that alternative methods outperformed SMO.

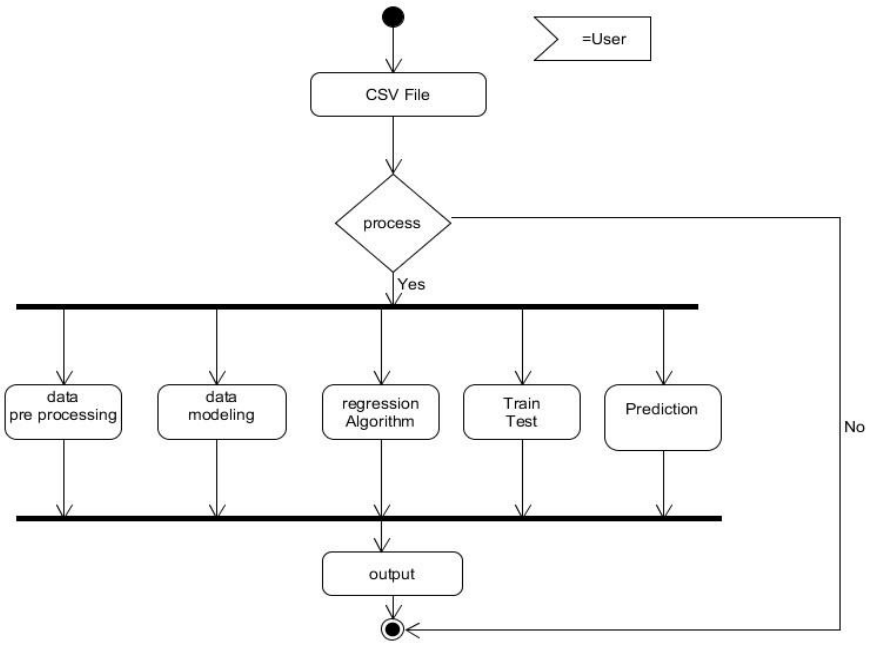
The paper highlights the uses of data mining methods, including techniques like K Means, K Nearest Neighbour, Artificial Neural Networks (ANN), and Support Vector Machines (SVM), for forecasting yield of crop. The aim is to identify effective data models that give best accuracy and generalization in yield predictions. The study evaluates various data mining techniques across different datasets.

The research seeks to develop predictive model that guides farmers in achieving high yield of crops through data mining techniques. Unlike traditional statistical approaches, data mining uncovers hidden insights through data analysis. Real data from farmers along the Thamirabarani river basin is utilized, with K-means clustering and decision tree classifiers applied to meteorological and agronomic data. Performance comparison indicates random forest algorithm outperforms other methods, and classifying clustered data yields high accuracy. The resulting rules assist farmers in making informed, proactive decisions before harvest.

close-range sensing methods offer the ability to survey crop and soil parameters that influence crop yield variations. When used with machine learning (ML) algorithms, these technologies can be leveraged to extract valuable information for managing yield of crop. The investigation employed four ML algorithms— linear regression (LR), elastic net (EN), k-nearest neighbours (k-NN), and support vector regression (SVR)—to forecast potato tuber yields using data gathered from soil and crop characteristics via close-range sensing. Data were elicited from 6 fields in Atlantic Canada, including three in Prince Edward Island and 3 in New Brunswick, across two growing seasons (2017 and 2018). The study highlighted the necessity of extensive datasets for robust model performance and emphasized the importance of developing site-specific management zones for potatoes to support global food security efforts.

Advantage:
The knn technique showed suboptimal performance in 3 out of four datasets—NB-2017, NB-2018, and PE-2017—resulting in root mean squared errors (RMSE).

III. METHODOLOGY



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13903

The dataset includes variables such as crop type, temperature, rainfall, pesticides usage, and yield data, collected from agricultural reports and surveys. This dataset is stored as a CSV file (yield_df.csv). Categorical variables like Item (crop types) are encoded using the Label Encoder to convert them into numerical form for use in machine learning models. The selected features for prediction include temperature, rainfall, crop type, and pesticide usage. The dataset is split into training and testing subsets using an 80/20 or 70/30 ratio for model evaluation. Correlations between features and the target variable (yield) are explored to identify influential factors. Various visualizations such as bar charts, scatter plots, and heatmaps are generated using Plotly to uncover relationships between features like rainfall, temperature, pesticides usage, and crop yield. Comparative visualizations across different areas and crops are also produced.

Multiple machine learning models are implemented, including:

- Random Forest Regressor
- Linear Regression
- Gradient Boosting Regressor
- XGBoost Regressor
- K-Nearest Neighbours (KNN)
- Decision Tree Regressor
- Bagging Regressor

The models are trained using the training dataset. Hyperparameters are set based on standard best practices for each algorithm. For example, n_estimators for Random Forest and max_depth for decision trees. Models are evaluated using accuracy scores, Mean Squared Error (MSE), and R-squared values. Comparisons between predicted and actual values are visualized using scatter plots. Random Forest and Gradient Boosting models provide feature importance scores, identifying the most influential factors contributing to crop yield. Insights derived from feature importance and visualizations are used to explain how temperature, rainfall, pesticides, and crop type impact agricultural productivity. The models are compared based on their performance metrics. The accuracy of each model is reported for both training and testing datasets. Comparative plots between actual and predicted values are generated for each model, along with tables showing performance metrics for easy comparison.

A user-friendly Streamlit interface allows users to input key variables such as temperature, rainfall, crop type, and pesticide usage. This triggers the prediction model to forecast the crop yield. Upon submission, the selected machine learning model predicts the crop yield, and the result is displayed within the web app, along with a formatted explanation of the input factors and predicted yield.

IV. ALGORITHMS

Random forest algorithm

The Random Forest algorithm was employed as the primary predictive model for estimating crop yields based on various environmental and agricultural factors. Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mean prediction of the individual trees. This technique is particularly robust to overfitting and handles high-dimensional data effectively.

The data was split into training and testing sets (e.g., 70% training, 30% testing). A Random Forest Regressor was trained on the training set, with the parameters chosen through experimentation to balance performance and generalizability. Key parameters included the number of trees in the forest and the maximum depth of the trees. After training, the model was evaluated on the test set using metrics such as mean squared error (MSE) and R^2 score to determine its accuracy in predicting crop yields. The Random Forest model demonstrated strong predictive performance, particularly in capturing the complex relationships between environmental factors and crop yield. In addition to Random Forest, other machine learning models such as Linear Regression, Gradient Boosting, XGBoost, and Decision Trees were tested. Random Forest provided a balanced performance with high accuracy and low mean squared error, outperforming many alternative models.

Linear Regression

Linear Regression is a fundamental statistical technique used to model the relationship between a dependent variable (crop yield) and one or more independent variables (such as temperature, rainfall, pesticides, and crop type). In this project, Linear Regression was applied to predict crop yield based on several factors affecting agricultural productivity. The Linear Regression model's performance was compared with more complex algorithms like Random Forest and Gradient Boosting.

DOI: 10.17148/IJARCCE.2024.13903

The model's accuracy on the test data was computed along with other performance metrics like Mean Squared Error (MSE) and R-squared (R^2). Although Linear Regression performed reasonably well, it was less effective in capturing the complex non-linear relationships between the features and the crop yield compared to Random Forest. For example, the R^2 value for Linear Regression was lower than that of Random Forest, indicating that Linear Regression had a limited ability to explain the variability in crop yield based on the input features. This suggests that more sophisticated models, such as Random Forest, are better suited for capturing the complexities of agricultural data.

Gradient boost regressor

Gradient Boosting Regressor is an advanced machine learning algorithm that builds an ensemble of decision trees to improve predictive accuracy. It iteratively combines weak learners, typically shallow decision trees, by correcting their previous mistakes to create a strong predictive model. The Gradient Boosting Regressor demonstrated improved predictive accuracy compared to simpler models like Linear Regression. The performance was evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R^2). The model's R^2 score on the test data was significantly higher, indicating that it was better able to capture the complex relationships between the features and the target variable. Additionally, the model excelled at handling non-linear relationships in the data, making it more suitable for capturing the impact of various environmental and agricultural factors on crop yield. For example, the Gradient Boosting Regressor was able to model the diminishing returns of yield increases despite high pesticide usage, a relationship that simpler models like Linear Regression could not fully capture.

However, the complexity of the Gradient Boosting Regressor comes with increased computational cost and a risk of overfitting if not properly tuned. Careful tuning of hyperparameters like the learning rate, number of estimators, and maximum depth was essential to prevent the model from memorizing the training data rather than generalizing well to unseen data.

XGBoost Regressor

XGBoost (extreme Gradient Boosting) is a highly efficient and scalable implementation of the gradient boosting framework. It improves upon the traditional Gradient Boosting Regressor by incorporating techniques like regularization, parallel computation, and advanced tree-pruning methods.

K-Nearest neighbours (KNN)

K-Nearest Neighbours (KNN) is a simple yet effective machine learning algorithm used for both classification and regression tasks. In the context of regression, KNN predicts the target value by averaging the values of the k nearest neighbours in the feature space. The K-Nearest Neighbours (KNN) Regressor provided a baseline for predicting crop yield, demonstrating its ability to model local patterns in the data. However, it generally performed less effectively than more sophisticated models like Random Forest and XGBoost, particularly in capturing complex interactions and non-linear relationships. Despite its limitations, KNN remains a valuable tool for its simplicity and ease of interpretation.

Decision Tree Regressor

The Decision Tree Regressor is a versatile and interpretable machine learning algorithm used for regression tasks. It models the target variable by constructing a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a predicted value

Bagging Regressor

Bagging (Bootstrap Aggregating) is an ensemble learning technique designed to improve the accuracy and stability of machine learning models. The Bagging Regressor, specifically, utilizes this approach for regression tasks by combining the predictions of multiple base regressors to enhance predictive performance and reduce variance. By aggregating predictions from multiple base regressors, bagging effectively reduced variance and provided more reliable crop yield predictions. Although it was generally less accurate than more advanced ensemble methods like Random Forest and XGBoost, it offered a valuable approach for enhancing the robustness and performance of regression models.

V. RESULT AND DISCUSSION

Each model's performance is evaluated using several metrics. Accuracy indicates how well the model generalizes to unseen data. Mean Squared Error (MSE) measures the average squared difference between predicted and actual values, with lower values signifying better performance. The R-squared (R^2) Score represents the proportion of variance explained by the model, where higher values denote a better fit. Visualization insights include the bar plot that displays the average yield per area, offering a clear view of which areas have the highest mean yields. Bar plots also show the average rainfall and pesticide usage per year by area, providing information on how these factors vary across different

DOI: 10.17148/IJARCCE.2024.13903

regions and their potential correlation with yield data. Scatter plots illustrate the relationships between yield and various features such as pesticides, rainfall, and temperature, helping to understand how these factors influence yield. The tabulated results highlight the most productive areas for each crop based on yield, which can guide agricultural practices and policy decisions.

VI. DISCUSSION

Random Forest and **XGBoost** consistently outperform other models, highlighting their effectiveness in capturing complex patterns and interactions in the data. These models are recommended for making accurate yield predictions.

Linear Regression serves as a reference point but may not capture non-linear relationships as effectively as ensemble methods or boosting algorithms. The performance metrics (accuracy, MSE, R2 score) indicate that the chosen models are well-suited for this prediction task, with Random Forest and XGBoost providing the best results.

The Streamlit application provides a user-friendly interface for crop yield prediction. Users can input:

- Average temperature
- Average rainfall
- Crop type
- Pesticides usage

The model generates predictions based on these inputs, providing a real-time estimate of crop yield. This interactive feature enhances usability and practical application of the model.

VII. CONCLUSION

The analysis highlighted the significant impact of average temperature, average rainfall, and pesticide usage on crop yield, providing valuable insights for agricultural planning. The Streamlit application effectively translates complex model predictions into a user-friendly format, allowing for real-time yield estimation based on user inputs.

The project's outcomes emphasize the potential of machine learning in enhancing agricultural productivity and decision-making. Future work could involve incorporating additional data features and refining the models to further improve prediction accuracy. Overall, this project provides a practical and valuable tool for optimizing crop yield through data-driven insights.

REFERENCES

[1]. O.D. Sirotenko and V.A. Romanenkov “Mathematical Models of Agricultural Supply” MATHEMATICAL MODELS OF LIFE SUPPORT SYSTEMS – Vol. II
[2]. J. Liu, X. Zhang, 2017: "Predicting crop yield with Random Forest in the context of agricultural decision-making." *Journal of Agricultural Informatics*, *8*(3), 21-30.
[3]. R. Patel, M. Desai, 2018: "Application of Random Forests for crop yield forecasting." *International Journal of Agricultural Sciences*
[4]. C. Philip Cox “A Simple Alternative To The Standard Statistical Model For The Analysis Of Field Experiments With Latin Square Designs”
[5]. Datasets from “Karnataka State Natural Disaster Monitoring Center” https://www.ksndmc.org/Weather_info.aspx
[6]. Datasets from “Directorate of Economics and Statistics” ANNUAL RAINFALL REPORT OF 2010.
[7]. Datasets from “Directorate of Economics and Statistics” ANNUAL RAINFALL REPORT OF 2011.atasets from “Directorate of Economics and Statistics” ANNUAL RAINFALL REPORT OF 2013.
[8]. Datasets from “Directorate of Economics and Statistics” ANNUAL RAINFALL REPORT OF 2014.
[9]. Datasets from “Directorate of Economics and Statistics” ANNUAL RAINFALL REPORT OF 2015.
[10]. Details regarding Crop and yield data from “JSS Krishi Vidya Kendra”, Suttur.
[11]. Jiawei Han, Micheline Kamber and Jian Pei “Data Mining – Concepts and Techniques” Third edition