RAJALAKSHMI ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

25 years OF ACADEMIC EXCELLENCE

Department of Computer Science and Engineering

CROP YIELD PREDICTION USING RANDOM FOREST REGRESSION

Mrs. Divya M,
Assistant Professor

220701209 Raghul A

Problem Statement and Motivation

- Predicting crop yield accurately is vital for agriculture, food security, and economic stability. Traditional methods are often inaccurate due to variable factors like weather and soil conditions. Climate change increases this uncertainty. Precise forecasts are needed for better planning, resource allocation, and risk management for farmers and supply chains. Developing a data-driven system to integrate diverse factors and improve yield prediction accuracy is a significant challenge and motivation for this project.

Existing System

- Current crop yield prediction methods include statistical models and machine learning techniques. Some use historical data; others add weather or satellite info. Many struggle to combine diverse factors like soil type, fertiliser use, and localised weather effectively. Simpler models may miss complex relationships. Advanced methods require extensive data. These limitations highlight the need for a more integrated approach that leverages various data sources for better prediction accuracy in diverse settings.

Objectives

- The project aims to accurately predict crop yield using machine learning. Key objectives are:
- 1) Collect and preprocess a dataset with agricultural, environmental, and management factors.
- 2) Train a robust Random Forest Regressor model.
- 3) Evaluate model performance using RMSE and R^2.
- 4) Identify important features influencing yield.
- 5) Build an interface for user input and prediction.

Abstract

- This project develops a machine learning model using Random Forest Regression for crop yield prediction. It incorporates factors like Region, Soil Type, Rainfall, and management practices. The dataset was preprocessed, including categorical encoding. The trained model achieved an RMSE of 0.5244 and an R^2 score of 0.9038, demonstrating strong predictive accuracy. Feature importance analysis revealed key yield drivers. This work provides a reliable model for improved crop yield forecasting.

Proposed System

- The proposed system predicts crop yield through a machine learning pipeline. It involves loading and preprocessing a dataset with diverse agricultural parameters. A Random Forest Regressor model is trained on this data. The model's performance is then evaluated using standard metrics. Finally, a user interface allows inputting specific conditions to obtain a predicted yield from the trained model.

System Architecture

- The system architecture consists of several layers:
- Data Input (raw dataset), Data Preprocessing (cleaning, encoding, splitting), Model Training (Random Forest Regressor training), Prediction & Evaluation (generating forecasts and calculating metrics)
- An User Interface in Colab for input and displaying predictions. Data flows from input through preprocessing to the trained model for prediction

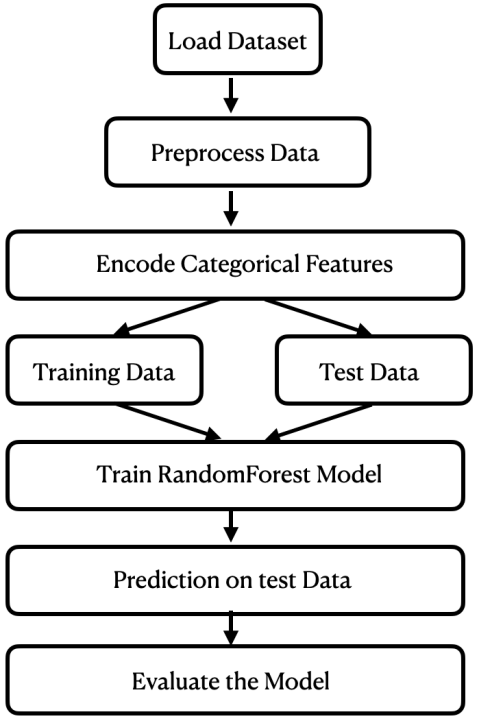
List of Modules

- **Dataset :** Loads and provides dataset crop_yield.csv (Kaggle).
- **Data Preprocessing:** Cleans data, encodes categories, splits train/test sets.
- **Model Training:** Trains the Random Forest Regressor.
- **Prediction and Evaluation:** Predicts on test data, calculates RMSE and R^2.
- **Input Prediction Interface:** Takes user input, preprocesses, gets prediction.

Functional Description for each modules with DFD and Activity Diagram

- **Dataset Description:** Reads the raw data file (CSV) into a DataFrame, providing information on columns and data types.
- **Data Preprocessing:** Transforms raw data; converts categorical text labels to numbers using Label Encoding; separates features and target; divides data into training and testing subsets.
- **Model Training:** Initialises the Random Forest Regressor model and uses the training data to fit the model, learning the relationships between input features and yield.
- **Prediction and Evaluation:** Uses the trained model to forecast yield on the test dataset and calculates the Root Mean Squared Error (RMSE) and R^2 score to quantify prediction accuracy.
- **Input Prediction Interface:** Prompts the user for feature values, applies the necessary preprocessing (encoding), feeds the processed input to the trained model, and outputs the resulting yield prediction.

Functional Description for each modules with DFD and Activity Diagram



Implementation & Results of Module

- The project was implemented in Python using Pandas and scikit-learn. A Random Forest Regressor was trained on the preprocessed dataset. Model evaluation on the test set yielded: RMSE: 0.524400333715013 and R^2: 0.9038127538443331. These results indicate that the model accurately predicts crop yield, with low average error and explaining over 90% of yield variance based on the input features.

Conclusion & Future Work

- The project successfully built a robust Random Forest model for crop yield prediction with high accuracy (R^2=0.9038, RMSE=0.5244).
- It effectively integrates various influencing factors.
- Future work could integrate the system with an IoT network in fields to collect real-time sensor data (soil, weather). This would allow incorporating live data streams for more dynamic and precise predictions, aiding timely farmer decisions and potentially optimising agricultural outcomes.

References

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R., Kim, S.H., & Ort, D.R. (2016). Random Forests for Global and Regional Crop Yield Predictions. PLoS ONE, 11(6), e0156571.

Patel, N., & Patel, D. (2025). Crop Yield Prediction Using Random Forest Algorithm and XGBoost Machine Learning Model. International Journal of Research and Innovation in Social Science (IJRISS), 9(4), 123-130.

Kumar, S., & Singh, V. (2023). Random forest algorithm use for crop recommendation. ITEGAM-JETIA, 9(43), 34-41. https://doi.org/10.5935/jetia.v9i43.906

Zhang, Y., Li, X., & Wang, J. (2023). Integrating random forest and crop modelling improves the crop yield prediction of winter wheat and oil seed rape. Frontiers in Remote Sensing, 4, 1010978.

Sharma, R., & Gupta, A. (2025). Smart Crop Prediction Using Random Forest and Machine Learning Models. SSRN Electronic Journal.

Thank You