Mitti Sathi
Your Soil, Your Guide

A project report submitted to ICT Academy of Kerala

in partial fulfillment of the requirements

for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

submitted by

Team 4

P A Gautham Nair

Afidh S Muhammed

ICT ACADEMY OF KERALA
THIRUVANANTHAPURAM, KERALA, INDIA
Nov 2022

List of Figures

List of Abbreviations

- **RF** - Random Forest
- **NLP** - Natural Language Processing
- **WSGI** - Web Server Gateway Interface
- **R^2** - R-squared (coefficient of determination)
- **API** - Application Programming Interface
- **ML** - Machine Learning

Table of Contents

Abstract

Many economies rely on farming, yet farmers often struggle to select the best crop for their land due to a lack of reliable, data-driven insights into soil and environmental conditions. This challenge frequently results in poor yields, wasted resources, and financial losses, particularly in regions where agriculture is essential for livelihoods. To address this, we created *Mitti Sathi*, an intelligent system that offers tailored crop recommendations based on soil and environmental factors, along with yield predictions for selected crops.

The system employs two machine learning models: a Random Forest Classifier to determine the ideal crop based on soil composition and a Random Forest Regressor to predict yield by considering factors like soil nutrients, temperature, and humidity. These models achieved impressive results, with the classifier attaining a test accuracy of 94.17% and the regressor achieving an R^2 score of 0.96. *Mitti Sathi* is accessible as a web application, hosted online via the Render platform and locally using Flask. By integrating traditional farming practices with modern data science, *Mitti Sathi* empowers farmers to make informed decisions, fostering sustainable agriculture and better harvests.

1. Problem Definition

1.1 Overview

Agriculture has been facing challenges that has affected sustainability and productivity. One of the major problems is the lack of data-driven tools that can help farmers choose the most suitable crops for their land. Shifting weather patterns, different types of soil, and lack of access to relevant information often force farmers to rely on outdated methods, which may not represent the best practices available.Such inefficiencies lead to wasted resources, environmental damage, and financial setbacks, especially in areas where agricultural production is a matter of survival. Farmers will, therefore, be able to make informed decisions by applying data science and technology to farming operations. This makes it possible for farmers to leverage technology and data science in agriculture to make informed decisions and address these inefficiencies.

1.2 Problem Statement

Agriculture is the backbone of many global economies, livelihoods, and food security. However, crop selection and yield prediction are critical decisions that many farmers make with little information, mostly relying on traditional knowledge and guesswork. This challenge arises primarily because of limited access to reliable, data-driven insights about their soil composition and environmental conditions. This leaves the farmer unable to determine which crops are best suited to his fields, leading to inefficient resource utilization, suboptimal yields, and economic losses.

Apart from economic problems, these inefficiencies lead to the wastage of considerable amounts of water, fertilizers, and labor, which further deteriorate agricultural ecosystems. This is particularly true in regions where agriculture forms a primary livelihood, and failure or success at harvest time may be the difference between food security and survival. It is a problem that is complicated by the fact that farmers lack easy access to analytical tools that help them use data on soils and environmental conditions to adopt more modern, sustainable methods.

Without smart systems to guide the farms, the decision-making is reactive and imprecise, which mostly leads to agricultural failures. Therefore, this poses an urgent call for a solution using modern techniques in data science to deliver actionable insights for farms to select appropriate crops and accurately predict yield potential. This helps promote the use of sustainable farmlands, productivity of farming, and reduced economic and environmental vulnerabilities related to agriculture.

2. Introduction

Farming plays a crucial role in supporting economies and livelihoods worldwide. Yet, farmers often run into obstacles that make their job tougher. While old-school farming methods have their merits, they often can't keep up with today's farming issues like changing weather patterns and different soil types. Many farmers find it hard to pick the right crops for their fields or guess how much they'll harvest without good tools or data-backed insights. These problems can lead to wasted supplies lower output, and money troubles in places where farming is the main way people make a living.

As data science and machine learning grow, farmers have a chance to change their decision-making process. These tools can look at data to find patterns and trends giving farmers solid advice that fits their needs. By combining tech with old-school know-how, farming can become more productive.

Our project, Mitti Sathi, wants to help farmers by offering a clever system to predict the best crop to grow based on soil and weather conditions. It also estimates how much a chosen crop might yield. We hope to give farmers the guts to make smart choices and use their resources well by putting this info on an easy-to-use website. By making this information accessible through a simple web application, we hope to give farmers the confidence to make informed decisions, optimize resources, and achieve better harvests. This project is a step towards combining traditional farming wisdom with modern technology to create a more sustainable future for agriculture.

3. Literature Survey

Machine learning , especially the Random Forest algorithm, has brought huge advancements in how crop classification and yield prediction are handled. This survey focuses on research and real-life applications that helped shape reliable solutions for these tasks.

The article on Medium about crop recommendation using Random Forest Classifier shows how analyzing soil and environmental data helps to find the most suitable crops for a region. Features like soil pH, rainfall, and temperature were highlighted as key factors for predictions. Random Forest's ability to rank these features makes it a great choice for giving accurate recommendations. This ensures farmers can make better decisions to boost productivity and sustainability in their practices.

A study published in MDPI explains how combining Random Forest with Particle Swarm Optimization (PSO) can improve classification accuracy, particularly for crop mapping. The hybrid model refines the parameters to improve predictions. It also stresses how using spatial data can improve agricultural applications, making this combination a valuable step forward in crop classification.

The Kaggle project on crop prediction demonstrates how Random Forest Classifier can be used for real-world data analysis. It focuses on cleaning and preparing data—like handling missing values and normalizing variables—before training the model. This process improves the model's reliability and performance. It also highlights the importance of making the model's results interpretable, which is essential for helping farmers understand and trust the predictions.

A study from ITM Conferences shows how combining Random Forest with Random Search for hyperparameter tuning improves yield prediction accuracy. It highlights how adjusting parameters correctly can boost the model's performance. The study also uses metrics like R^2 to validate the model, showing how essential it is to test models properly for real-world use cases.

Research study on cotton yield prediction shows how climate factors, like temperature and rainfall, affect agricultural outcomes. The study discusses using Random Forest to analyze these variables and predict yields. It stresses the need for datasets with enough detail to account for climate impacts. This research sheds light on how tools like Random Forest can help farmers deal with climate variability in yield predictions.

A global study on crop yield prediction focuses on Random Forest for multivariate regression tasks. The algorithm achieved a high R^2 score, which shows how well it can model complex relationships in agricultural data. The study also talks about the importance of preprocessing, like scaling data and managing missing values, which is critical for building solid regression models that work across different regions.

The preprocessing techniques used in this project, like MinMax and Standard scaling, was inspired by insights from studies that stressed the role of normalization and scaling in improving model performance. Also, methods such as synthetic data generation and unique outlier handling comes from the need to tackle dataset limitations and ensure robustness, as shown in research on climate variability and crop yield prediction.

The studies reviewed here show how powerful machine learning, especially Random Forest, can be in solving agricultural challenges. Random Forest's ability to handle complex data, combined with preprocessing and optimization methods, makes it an excellent tool for both crop classification and yield prediction.

For classification, the research emphasizes the need for good feature selection and making sure outputs are easy to interpret so farmers can use them. For regression, it's clear that using diverse features and fine-tuning the model are necessary for accurate yield predictions. Together, these approaches show how data-driven methods can improve farming efficiency and promote sustainable practices .

4. Result

4.1 Crop Type Classification (Classifier)

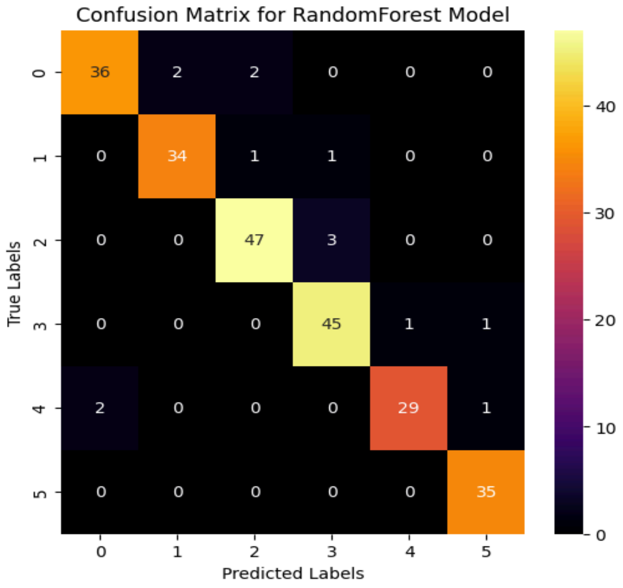

Figure 4.1.1

- **Confusion Matrix:**
  The Random Forest Classifier shows excellent predictive performance, with most predictions lying on the diagonal. Misclassifications are minimal, primarily occurring between closely related crop classes (e.g., Class 4 misclassified as Class 5).

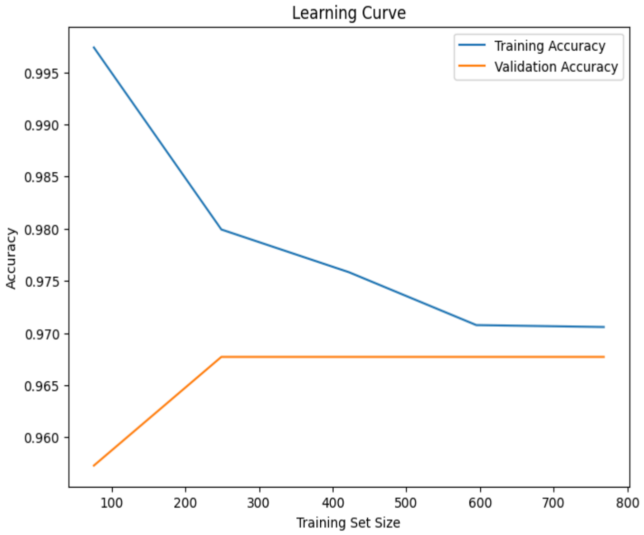

Figure 4.1.2

- **Learning Curve:**
  The classifier achieves high accuracy, with validation accuracy plateauing after 600-800 samples. While a slight gap exists between training and validation curves, this indicates mild overfitting, but overall, the model generalizes well.

- **Model Accuracy:**
  - Training Accuracy: 97.19%
  - Test Accuracy: 94.17%

4.2 Crop Yield Prediction (Regressor)

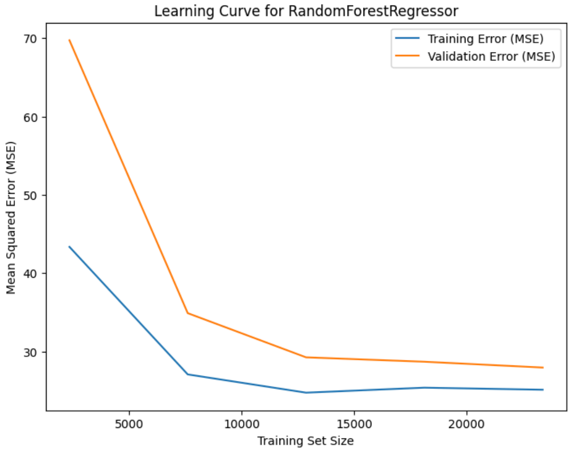

Figure 4.2.1

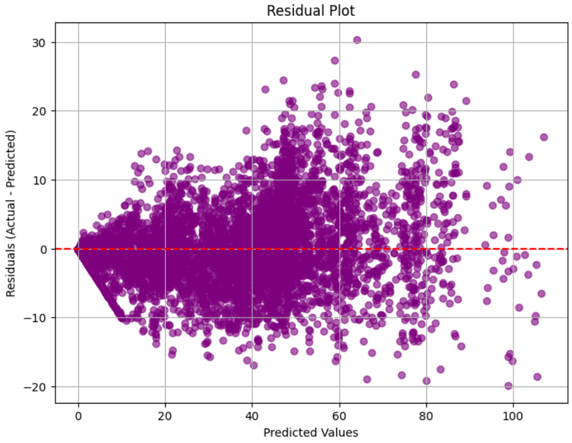

Figure 4.2.2

- **Learning Curve**:
  The Random Forest Regressor shows strong generalization as training error stabilizes at a low value, and validation error converges towards it. This demonstrates the model's ability to predict crop yield effectively based on features like soil NPK levels, temperature, and humidity.

- **Residual Plot**:
  The residuals are symmetrically distributed around zero, indicating unbiased predictions. While slight variability is observed at extreme yield values, the absence of patterns in the residuals confirms the model effectively captures the nonlinear relationships between the input features and crop yield.

- **Model Accuracy**:
  - Training $R^2 = 0.96$
  - Test $R^2 = 0.96$
  - Test Accuracy within ±10% tolerance: 32.20%

4.3 Model Deployment

The deployment of the crop classification and yield prediction models involved using Flask, Render, Waitress, and standard web development tools to ensure both accessibility and a user-friendly interface at MittiSathi.com . Flask, a lightweight Python framework, was utilized to host the models locally, managing HTTP requests and responses, loading pre-trained models (pickle files), and returning predictions. Static HTML pages were served directly, providing a simple and effective user interface for local testing. For web deployment, Render was chosen with the codebase pushed to GitHub and linked to Render, which automated the process. This allowed the application to be accessible online via a stable web URL for remote users. Waitress was employed as the WSGI server to ensure the application was production-ready, efficiently handling concurrent requests and improving performance compared to Flask's built-in development server.

The frontend of the application was built using HTML for structure, CSS for styling, and JavaScript for interactivity. These static web pages enabled users to input data and view predictions seamlessly, offering a clean and responsive interface. This deployment strategy combined a lightweight backend, scalable hosting, and responsive frontend design to create a functional and reliable application for users.

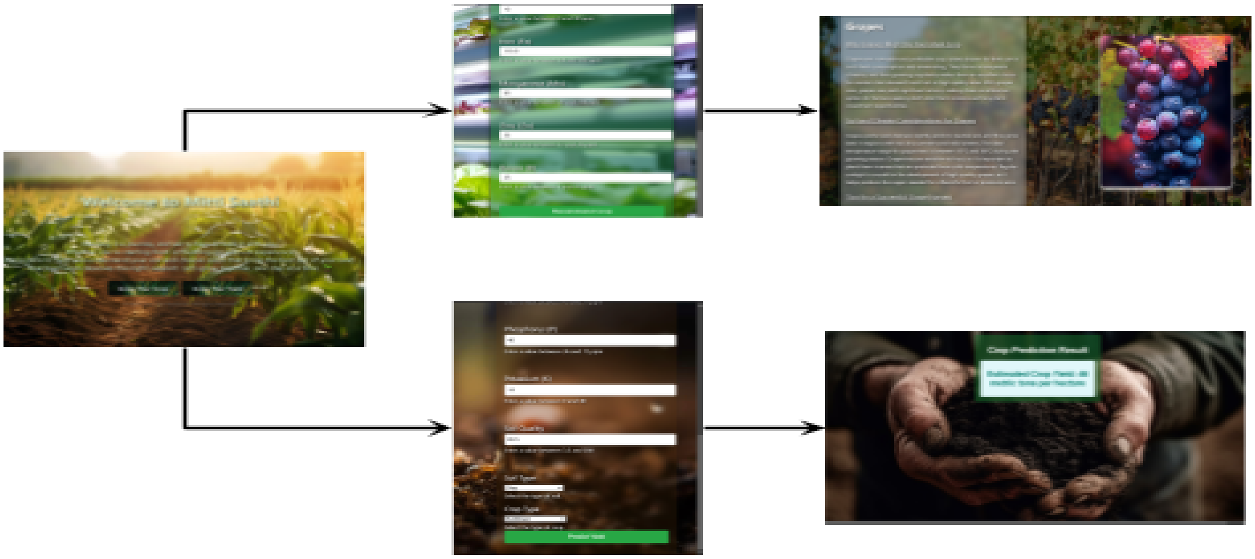

Figure 4.3.1

5. Conclusion

In conclusion, the deployment of the crop classification and yield prediction models has shown the potential of using machine learning to improve agricultural decision-making. By integrating Flask for local hosting, Render for web deployment, Waitress for efficient handling of requests, and web development tools for an intuitive user interface, the application is now accessible to users both locally and online. This solution empowers farmers with data-driven insights, giving them a more informed approach to crop selection and yield prediction.

Looking towards the future, there’s a lot of potential to enhance this project even further. A more specialized focus on region-specific crops and environmental conditions will allow for more precise and tailored recommendations based on local soil and climate data. This would make the system even more valuable for farmers in different areas, providing them with more accurate and actionable cultivation advice. Also, adding natural language processing (NLP) to the interface could help make the tool more accessible to a wider range of users to communicate and interact with the system with more ease and in their own language . This development would ensure that the application becomes even more user-friendly, helping farmers adopt data-driven, sustainable practices more easily.

6. References

1. **Optimal Crop Recommendation**
   Medium Article: "Optimal Crop Recommendation Using a Random Forest Classifier."
   Published by Insights of Nature
   URL :
   https://medium.com/insights-of-nature/optimal-crop-recommendation-using-a-random-forest-classifier-e2de0b77c7f7

2. **Crop Mapping with RF and Particle Swarm Optimization**
   Yan, W., et al. (2020). "Improving Crop Mapping Accuracy Using Random Forests and Particle Swarm Optimization."
   Remote Sensing, MDPI. Vol. 12, Issue 9, Article 1449.
   URL : https://www.mdpi.com/2072-4292/12/9/1449

3. **Crop Prediction Using RF Classifier**
   Kaggle Project: "Crop Prediction Using Random Forest Classifier."
   By Karthik Reddy.
   URL :
   https://www.kaggle.com/code/karthikreddy77/crop-prediction-using-random-forest-classifier

4. **Improved Yield Prediction**
   Al-Ayyoub, M., et al. (2023). "Hybrid Machine Learning for Improved Crop Yield Prediction Using Random Search and Random Forest."
   ITM Web of Conferences. Vol. 44, Article 02007.
   URL :
   https://www.itm-conferences.org/articles/itmconf/pdf/2023/06/itmconf_icdsac2023_02007.pdf

5. **Cotton Yield Prediction**
   Xu, L., et al. (2023). "Machine Learning for Cotton Yield Prediction under Climate Variability."
   arXiv preprint.
   URL : https://arxiv.org/abs/2312.02299

6. **Multivariate Regression Study**
   Ahmed, S., et al. (2023). "A Multivariate Machine Learning Framework for Global Crop Yield Prediction."
   arXiv preprint.
   URL : https://arxiv.org/abs/2312.02254