CSE 343: Machine Learning Project Proposal

Naman Jindal, naman22311@iiitd.ac.in

Nishchay Sharma, nishchay22331@iiitd.ac.in

Harshil Handoo, harshil22206@iiitd.ac.in

Himanshu Kumar, himanshu22215@iiitd.ac.in

AgriPredict: Machine Learning-Based Crop Yield Prediction

In an era where agricultural efficiency and sustainability are more critical than ever, accurate crop yield prediction has become a key factor in ensuring food security and optimizing resource management. **AgriPredict leverages machine learning, by analyzing a range of influential factors such as area, rainfall, wind, crop type, and soil condition,it aims to deliver precise and actionable insights, predicting the quantity of crop yield.** This empowers farmers and agricultural stakeholders to make informed decisions and maximize productivity.

1. Motivation

1.1. Why this project?

Agriculture plays a crucial role in global food security, yet predicting crop yield remains challenging due to complex interactions between environmental factors. Traditional methods often rely on historical data and intuition, which may not fully capture modern agriculture's dynamic nature. Advanced, data-driven approaches are essential for improving crop yield predictions, especially in the face of climate change and resource constraints. This project aims to address this gap by applying machine learning techniques to enhance agricultural productivity and sustainability.

1.2. How did we think about this?

The idea for AgriPredict was driven by the need to ensure food security amid climate challenges and population growth. We recognized that accurate crop yield prediction can significantly improve farming practices by enabling informed decision-making. Leveraging machine learning to analyze various factors affecting yield, our goal is to provide insights that help farmers optimize their practices and enhance productivity, contributing to a more secure and sustainable food supply.

2. Related Work

2.1. Crop Yield Prediction Using Machine Learning: A Pragmatic Approach

This study focuses on improving crop yield prediction in India's agricultural sector using machine learning techniques such as Random Forest, Adaboost, Gradient Boost, and SVM. It finds that Random Forest achieves the highest accuracy, helping farmers enhance their yields.

2.2. Yield Prediction for Crops by Gradient-Based Algorithms

This study evaluates machine learning algorithms for predicting crop yields, finding that CatBoost, LightGBM, and XGBoost outperform others.

3. Timeline

In the first week, review the Kaggle dataset, set project goals, and set up the environment. Week two involves EDA, data cleaning, and splitting the dataset. Focus on feature engineering in week three. Train models (Linear Regression, Random Forest, Gradient Boost, XGBoost, KNN, Decision Tree) in week four. Evaluate model performance and fine-tune in week five. Perform advanced training and comparison in week six. Optimize and validate the best model in week seven. Prepare for deployment and create necessary documentation in week eight. Deploy and monitor the model in week nine, and review the project, finalize the report, and make any necessary adjustments in week ten.

4. Individual Contributions

- **Naman**: Responsible for model building and evaluation, as well as data collection and preparation.

- **Nishchay**: Focuses on model building and evaluation, with an emphasis on exploratory data analysis (EDA).

- **Harshil**: Handles data collection and preparation, feature engineering, and model optimization.

- **Himanshu**: Works on model optimization and EDA, as well as setting up working environments.

5. Final Outcome

The primary goal of this project is to develop a robust machine learning model that accurately predicts crop yield based on multiple factors. We aim to contribute to agricultural data science by creating a model that assists farmers, policymakers, and researchers in making informed decisions about crop production and resource allocation. The final deliverable will include a trained model, a detailed report on the findings, and a comprehensive evaluation of the model's performance.