Weekly Progress Report

Name: Md khurshid alam
Domain: Machine Learning
Date of submission:02/11/2025

Final Report: Crop

I. Overview:

This week, the primary focus was on understanding the agricultural datasets related to crop production in India and preparing the data pipeline for model development.
The goal of this project is to analyze and predict agricultural crop production using historical data from multiple sources covering crop types, cost of cultivation, area, yield, seasonal data, and government indices.

A total of five CSV files were analyzed and integrated, covering:

- Crop-wise cultivation cost and yield by state
- Production, area, and yield across 2006–2011
- Crop varieties, season duration, and recommended cultivation zones
- Agricultural index values across years
- National agricultural production time-series data (1993–2014)

This week involved studying the data, cleaning formats, handling year patterns, merging datasets, and creating a complete Python pipeline to prepare and process the data for machine learning.

II. Achievements:

Dataset Understanding

- Reviewed the structure and content of all five agricultural datasets
- Identified key features such as Crop, State, Year, Production, Area, Yield, Cost metrics, and Variety details
- Noted varying year formats (e.g., 2006-07, 3-1993) and multiple unit formats requiring transformation

Data Pipeline Development

- Designed a unified pipeline to:
- Load CSVs
- Clean numeric formats (comma separators, units, mixed formats)
- Parse complex year formats
- Convert wide-format production tables into long format
- Merge datasets into a master table using Crop, State, and Year keys

Data Cleaning & Feature Engineering

- Converted mixed numeric fields (e.g., 12,34.56) into float values
- Filled missing production values using Area × Yield logic where applicable
- Categorical encoding for Crop and State variables
- Extracted and aligned time-series data with crop-specific production data

ML Model Baseline

- Implemented a baseline Random Forest Regressor
- Split data into train-test sets
- Generated evaluation metrics (MAE, RMSE, R^2)
- Saved model and prediction results for further analysis

III. Challenges

Challenge	Notes
Complex year formats	Required custom parsing for patterns like 2006-07, 3-1993, etc.
Non-uniform CSV structures	Each dataset followed a different schema, requiring careful merging logic
Missing & inconsistent numeric formats	Corrected thousands separators, decimal formats, and empty values
Limited overlapping fields	Some variety/state info had to be merged at crop-level only

IV. Learning Resources:

- Pandas documentation (pivoting, merging, string cleaning)
- Scikit-Learn documentation (Random Forest, train-test split)
- Tutorials on handling multi-source time-series agricultural datasets
- Articles on agriculture analytics & crop yield prediction models

V. Key Results & Insights:

1.Data Successfully Combined

All five agriculture datasets were cleaned and merged into one usable dataset.

2.Production Trend Increasing

Agriculture production in India generally increased over the years, with temporary drops in drought years (e.g., 2009–10).

3.Yield Is the Main Driver

Yield per hectare showed the strongest impact on production, followed by area cultivated.

4.Costs vs Output Insight

Higher cultivation cost does not always mean higher production; regional and seasonal factors play a major role.

5.Model Performance

The baseline Random Forest model achieved good accuracy, showing that crop production trends are learnable and predictable.

6. Improvement Areas

Some missing values and unmatched year formats remain; external data like rainfall and soil info can further improve predictions.

VI. Conclusion:

This week successfully established the foundation for the Agriculture Crop Production Prediction project. The datasets were explored, cleaned, and combined into a single structured format suitable for analysis. Initial modeling using a baseline Random Forest demonstrated promising predictive capability, confirming that historical agricultural trends in India are learnable and can support future production forecasting.

The project is now positioned to move into deeper feature engineering, model optimization, and detailed evaluation. With further enhancements, this system can help analyze crop patterns, support planning decisions, and provide insights beneficial for agricultural development and resource management.