Business Analytics for Managerial Decisions

Project Report

Submitted by:

Group 11

Shah Karan Bhavesh	**MBA/0065/61**
Heet Jain	**MBA/0105/61**
Devraj Shah	**MBA/0424/61**
Tanushri Tripathi	**MBA/0473/61**
Krushn Pathak	**MBA/0439/61**

1. Business Context

- **Challenges for the Government Due to Adverse Climate Conditions**
  In India, 80% of marginal farmers, farming less than one hectare and producing 60% of the country's agricultural output are increasingly vulnerable to droughts, floods and other adverse climatic events. These conditions cause 32-39% variability in crop yields, creating significant uncertainty in national food production forecasts.

- **Economic & Policy Pressures**
  Agriculture contributes 20% to India's GDP, yet climate-driven yield losses (estimated at 3.1% per °C rise) threaten to undermine growth targets. Unpredictable yields complicate the government's ability to plan food procurement, manage buffer stocks, control inflation, and allocate subsidies effectively. This volatility also forces reactive measures such as unplanned food imports, increased price support, and emergency relief spending, straining fiscal resources.

- **Need for Data-Driven Forecasting**
  For the government, accurate crop yield prediction using machine learning can strengthen policy-making, improve allocation of subsidies and disaster relief, stabilize market prices, and support climate-resilient agricultural programs. Reliable forecasts would also aid in long-term planning for food security, trade strategy, and risk mitigation for vulnerable farming communities

2. Problem Statement

Previously, we had thought about predicting crop yield through rainfall data but post our discussion we decided upon building a **crop yield prediction for the government** with a focus on the **macro-economic irrigation policies** & how they affect the crop yield data **in each state** & across the **Rabi & Kharif seasons**

3. Solution

- **Downloading Datasets**

The first stage involved collecting the necessary datasets for the analysis. Year-wise data was downloaded from the official [DESA Agriculture portal]

Four key datasets were used:

- **Classification of Area:** Provides detailed land classification categories (e.g., forest area, cultivable land, fallow land, net sown area) for each year.
  Link: https://data.desagri.gov.in/weblus/classification-of-area-report-web

- **Source Irrigated Area:** Breaks down irrigated land by source (e.g., canal, well, tube-well, tank, other sources) to understand changes in irrigation dependency.
  Link: https://data.desagri.gov.in/weblus/lus-source-irrigated-area-report-web

- **Crop Irrigated Area:** Provides crop-wise irrigation data, useful for identifying crop-specific water usage patterns.
  Link: https://data.desagri.gov.in/weblus/lus-crop-irrigated-area-report-web

– **Area under crops:** Contains annual statistics on the total area cultivated under different crops, enabling yield and productivity comparisons.
Link: https://data.desagri.gov.in/weblus/lus-area-under-crops-report-web

All datasets were reviewed for completeness, downloaded in a uniform year-wise format, and prepared for subsequent cleaning, transformation, and modelling steps.

4. Data Pre-Processing

Following were the steps taken for data pre-processing which included merging the datasets and data cleaning:

- **Merged Files:** Combined all four datasets (Classification of Area, Source Irrigated Area, Crop Irrigated Area, and Area under Crops) into a unified master file
- **Resolved Errors in Merging:** Addressed mismatches in keys (e.g., spelling variations) that caused join errors. Ensured that all district & year combinations were correctly aligned
- **Mapped Districts to States:** Mapped each district correctly to its respective state
- **Resolved Name Ambiguities:** Cleaned & standardized names across datasets to maintain uniformity
- **Dropped Redundant Columns:** Removed unnecessary columns
- **Handled Missing Values:**
    - Removed rows that were entirely empty.
    - Replaced partial missing entries with 0 where applicable (e.g., if irrigation data for a specific crop was missing but the rest of the row was valid)

5. Data Visualization

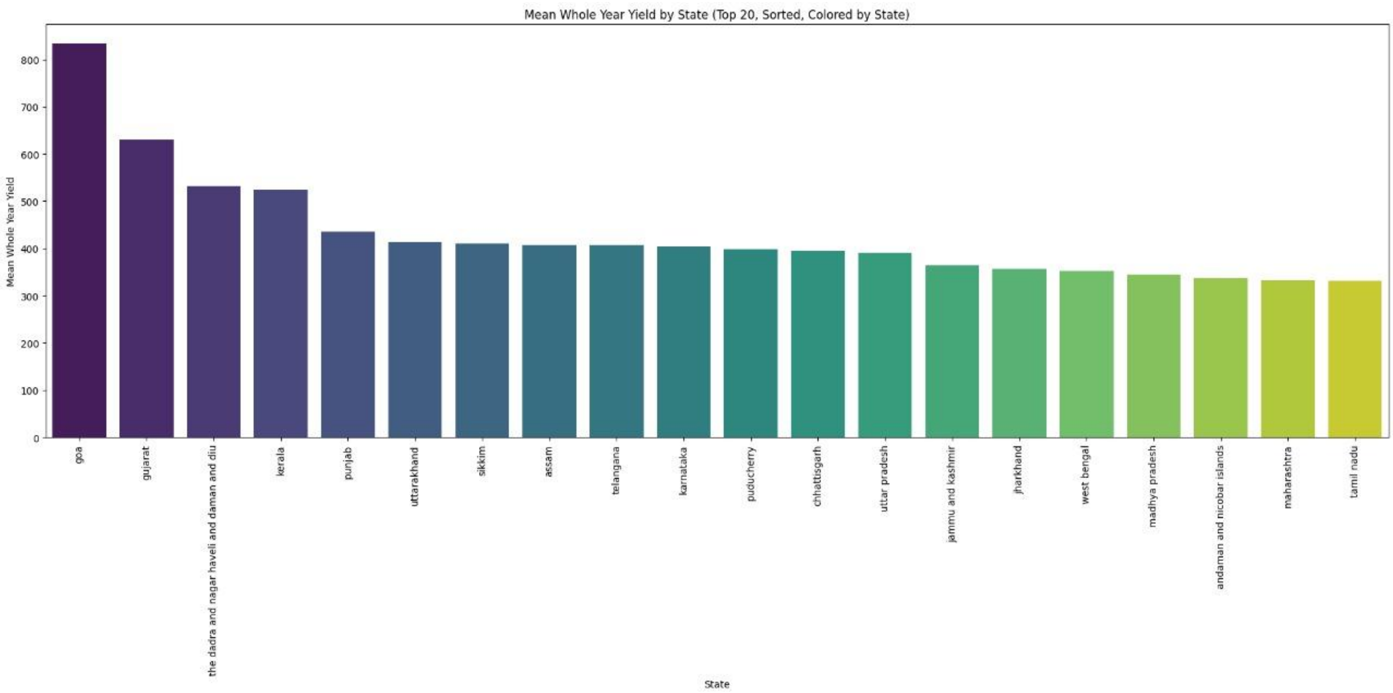

Exhibit 1: Average Yield for Top 20 States

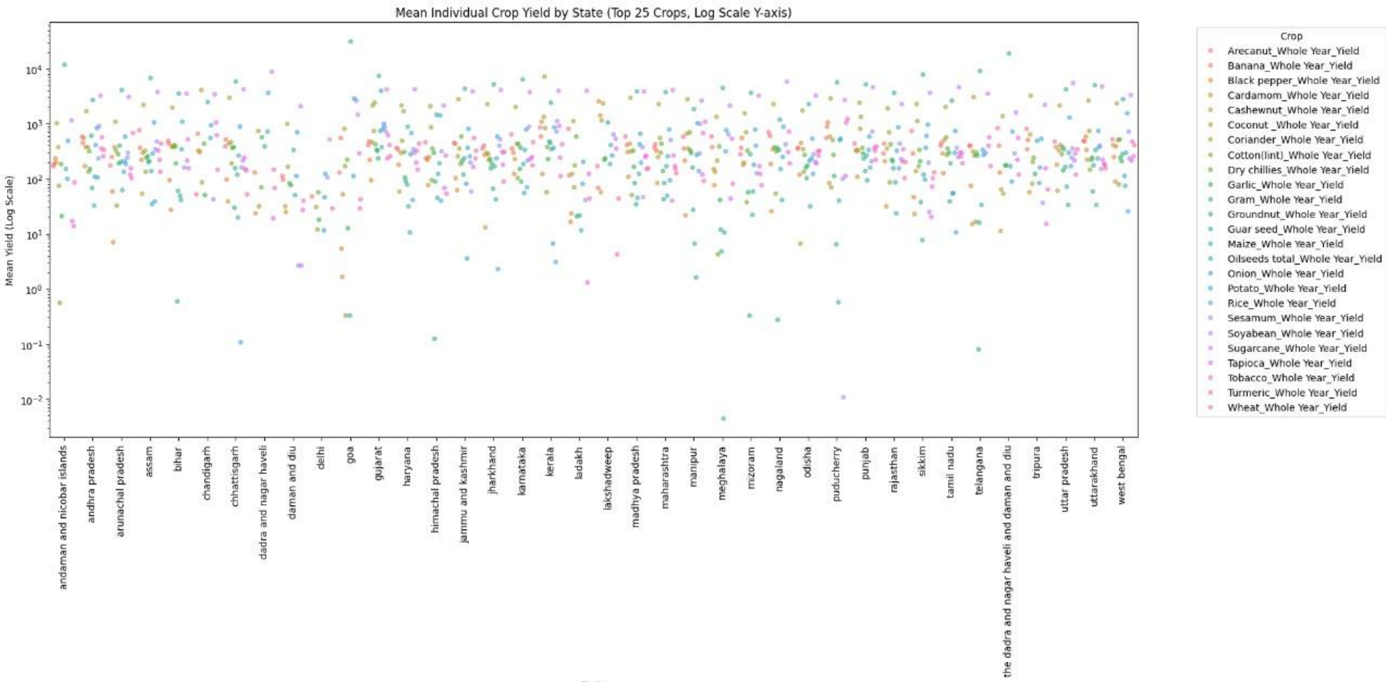

Exhibit 2: Average Yield Per Crop for Top 20 States

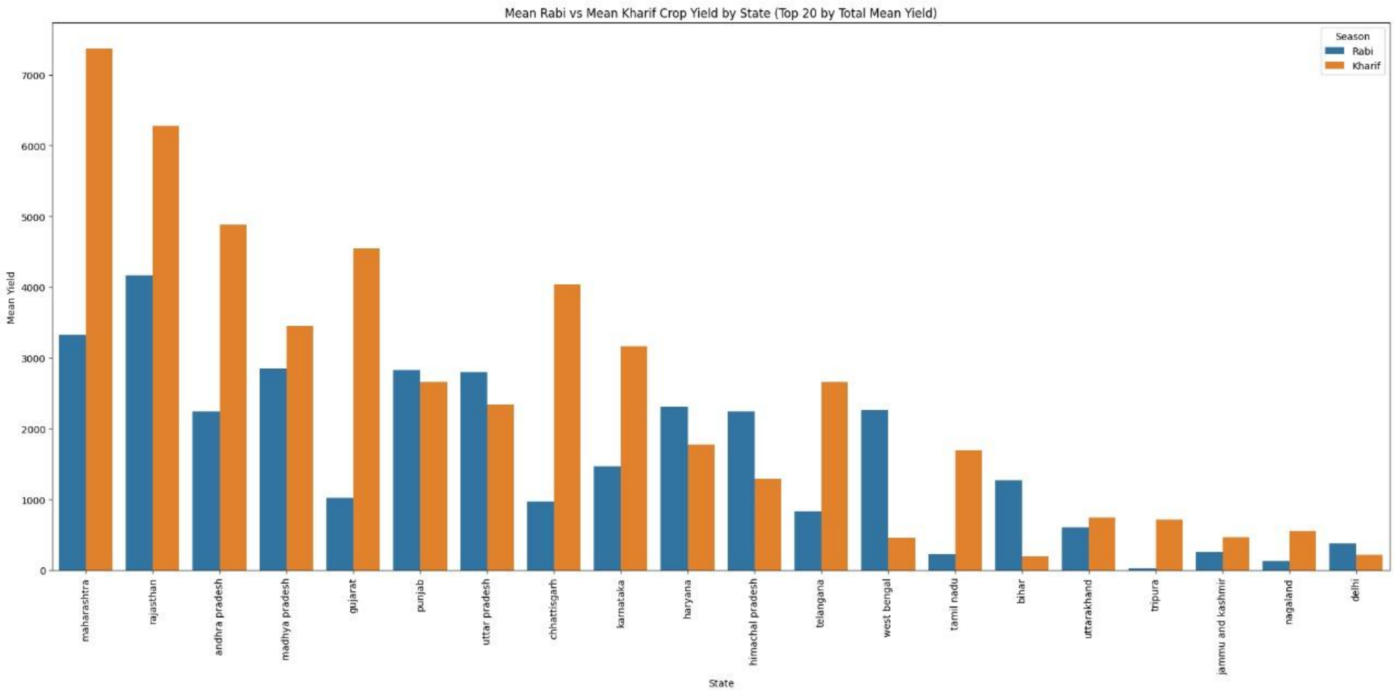

Exhibit 3: Mean Rabi & Kharif Crop Yield for Top 20 states

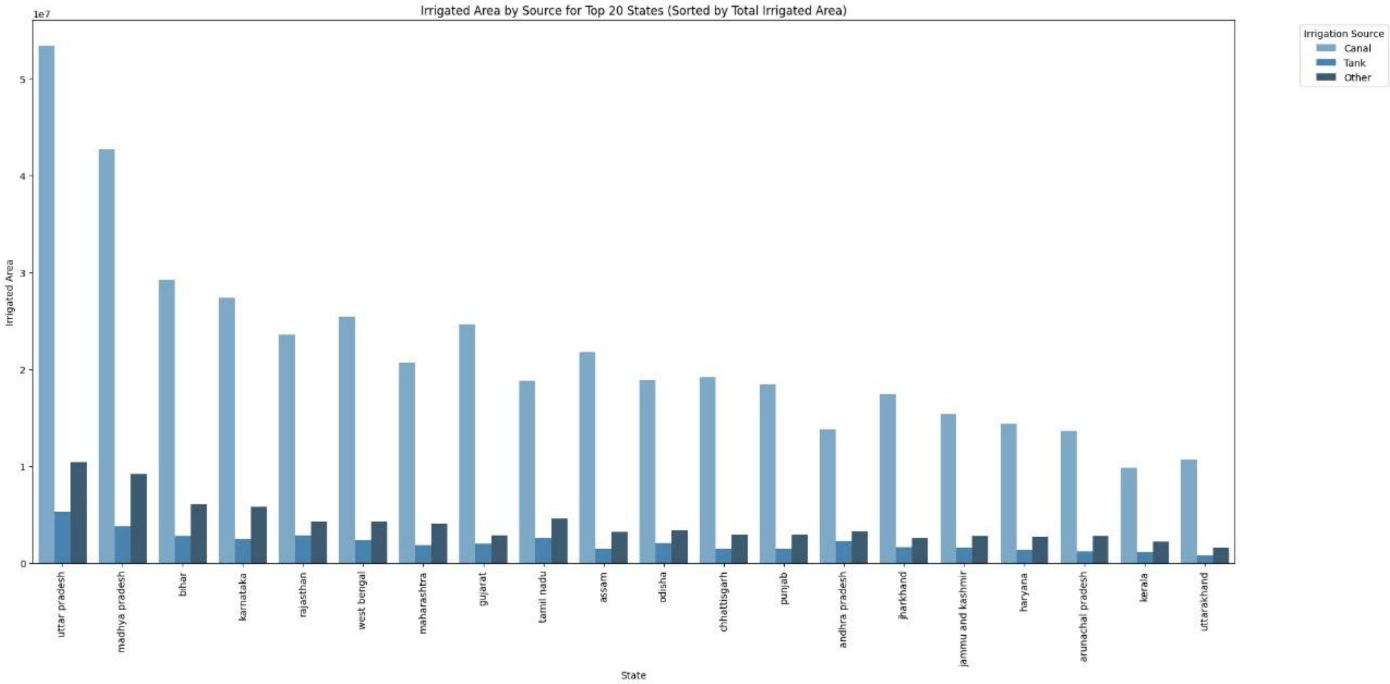

Exhibit 4: Irrigated Area By Source for Top 20 states

6. Prediction Models

We split the cleaned dataset chronologically into training and test sets using an 80:20 ratio for our prediction models

Models Used
Five different predictive models were implemented to compare performance:

- **Linear Regression:** *Average R^2 = 0.29*
- **Decision Tree Regressor (with hyperparameter tuning):** *Average R^2 = 0.48*
- **Random Forest Regressor:** *Average R^2 = 0.61*
- **Ensemble Learning with Soft Voting & the above 3 models:** *Average R^2 = 0.67*
- **Neural Network:** *Average R^2 = 0.13*

Model Evaluation

- We used **the average R^2 score** of all crops given by each model and the Mean Absolute Error as our evaluation metrics

Outcome
The Ensemble Learning approach with soft voting outperformed all other models in terms of average R^2 and was therefore selected as the final prediction model for our solution in the Streamlit App

7. Snippet of StreamLit App

Link to Streamlit App (Deployed): https://cropyield-hqpxjgfee4ejmzf6brcbx2.streamlit.app/
Link to Github: https://github.com/heetj2026/Business-Analytics-Project-Group-11

Thank You!