CAPSTONE PROJECT ON

“Agri Data Scope: Big Data Analytics for Sustainable Agriculture and Crop Intelligence”

Submitted in the partial fulfilment of the requirements for the award of the degree of

MASTER OF BUSINESS ADMINISTRATION

SCHOOL OF MANAGEMENT STUDIES

UNIVERSITY OF HYDERABAD

By

KYATHAM SRILAYA

24MBMA63

MBA (General) 2024–2026

Under the esteemed guidance of

SREE LAKSHMI MAM

DECLARATION

I, *Kyatham Srilaya*, hereby declare that the Capstone Project entitled **"Agri Data Scope: Big Data Analytics for Sustainable Agriculture and Crop Intelligence"**, submitted to the School of Management Studies, University of Hyderabad, in partial fulfilment of the requirements for the award of the degree of **Master of Business Administration,** is the result of my own work carried out under the guidance of **Sree Lakshmi Mam**.

The work embodied in this project is original and has not been submitted to any other institution for any degree, diploma, or fellowship. All references and data sources have been duly acknowledged.

Date: 11/11/2025

Place: School of Management Studies

University of Hyderabad

Signature of the Student

Kyatham Srilaya

Roll No: 24MBMA63

School of Management Studies

University of Hyderabad

Table of Contents

Chapter 1: Introduction

Agriculture remains the backbone of most developing economies, serving as a key driver for food security, employment, and GDP. However, it faces serious challenges such as climate change, water scarcity, soil degradation, and unpredictable rainfall patterns. Traditional farming methods relying on manual decisions and intuition are no longer sufficient to ensure sustainable productivity.

With the rise of Big Data Analytics, it has become possible to collect, process, and analyse vast volumes of agricultural data — from soil composition and weather conditions to rainfall, crop yield, and market data — to make intelligent decisions. This data-driven approach enables farmers and policymakers to improve productivity, manage resources effectively, and adapt to climatic changes.

The project titled "Agri Data Scope: Big Data Analytics for Sustainable Agriculture and Crop Intelligence" applies Big Data frameworks like Apache Spark and Databricks to design and implement an agricultural data pipeline. The goal is to demonstrate how large-scale agricultural datasets can be processed and analysed to extract meaningful insights.

- **Objectives of the Study**

1. Integrate agricultural datasets (FAOSTAT, IMD, ICAR, and climate data) using PySpark.
2. Clean, transform, and analyse data for better decision-making.
3. Implement five real-world use cases: yield prediction, rainfall impact, climate matching, soil recommendation, and GDD analysis.
4. Visualize insights through Databricks dashboards.
5. Demonstrate how Big Data tools enhance sustainable agricultural practices.

- **Need for the Study**

Massive data from sensors, satellites, and weather stations is underutilized in agriculture. Using Big Data tools, policymakers can identify optimal crop patterns, predict yield, manage irrigation and fertilizer, and adapt to environmental changes.

- **Scope and Significance**

This study focuses on integrating diverse agricultural datasets and applying predictive models to extract actionable insights. It demonstrates precision agriculture and resource optimization through advanced data analytics on the Databricks platform.

Chapter 2: Literature Review

- **Big Data in Agriculture**

Big Data Analytics transforms agriculture by enabling real-time decision-making, predictive modelling, and data-driven resource management. According to Wolfert et al. (2017), agricultural data originates from genomics, sensors, remote sensing, markets, and climate monitoring. Big Data tools such as Apache Spark and Databricks allow for scalable, distributed data processing.

- **Climate and Yield Analytics**

Climate strongly influences agricultural productivity. Ray et al. (2019) found that over 60% of crop yield variability in India is linked to rainfall and temperature. Machine learning models such as Linear Regression, Random Forest, and Decision Trees have proven effective in quantifying these effects. Tripathi et al. (2020) built weather-based yield models achieving $R^2 > 0.8$.

- **Rainfall Impact Studies**

Rainfall directly affects crop growth and harvest timing. Reddy & Sharma (2021) showed that inconsistent rainfall results in 10–20% yield loss. Big Data allows integrating rainfall with soil moisture and yield datasets, improving accuracy in correlation models.

- **Soil and Crop Suitability**

Soil health influences nutrient availability and productivity. The Soil Health Card program in India (ICAR) provides parameters like pH, nitrogen, phosphorus, and organic carbon. Joshi & Patel (2020) applied Decision Trees to recommend crop-soil alignment. The present project extends this by using machine learning classifiers for soil–crop recommendations.

- **Yield Prediction Models**

Predicting yield helps plan production and market strategies. Bendre & Thool (2016) used regression models to forecast yield, while Zhou et al. (2019) used Spark-based regression for large-scale forecasting. This project integrates historical yield and climate data using PySpark MLlib for scalable predictions.

- **Research Gap**

Although prior studies explored yield forecasting and climate correlations individually, few have combined multiple agricultural datasets into a unified Big Data architecture. The Agri Data Scope project bridges this gap by integrating multi-source datasets for comprehensive agricultural intelligence.

Chapter 3: Research Methodology

This study adopts a quantitative, exploratory, and analytical design, utilizing the Big Data Analytics Lifecycle: data collection, transformation, modelling, and visualization.

Data Sources

Dataset	Source	Description
FAOSTAT	FAO	Historical crop yield data (2000–2025)
Rainfall	IMD / Kaggle	Monthly and annual rainfall data
Soil	ICAR / SHC	pH, N, P, K, organic carbon
Climate	NOAA	Temperature, humidity, rainfall records
Crop Recommendation	Kaggle	Soil and climate suitability data

Tools and Technologies

- Apache Spark (PySpark): Distributed data processing
- Databricks: Cloud environment for pipeline creation
- Spark MLlib: Machine learning for prediction/classification
- Delta Lake: Reliable data storage and versioning
- Python (NumPy, Pandas): Data wrangling and preprocessing
- Matplotlib/Seaborn: Visualization tools

Pipeline Architecture (Medallion Model)

1. Bronze Layer – Raw Data: Import raw CSVs, standardize columns, and ingest into Databricks.
2. Silver Layer – Cleaned Data: Remove duplicates, handle missing values, and convert datatypes.
3. Feature Layer: Combine datasets and create derived variables like rainfall deviation and soil fertility index.
4. Model Layer: Implement ML models for each use case.
5. Visualization Layer: Use Databricks dashboards to present insights.

Model Techniques

- Linear Regression – Yield Prediction
- Correlation Analysis – Rainfall Impact
- K-Means Clustering – Climate Matching
- Decision Tree Classifier – Soil Recommendation
- GDD Calculation – Optimal crop growth periods

This multi-layer pipeline enables data-driven insights for sustainable agriculture.

Chapter 4: Results and Discussion

- **Dashboard Overview**

A unified Agriculture Insights Dashboard was created in Databricks to visualize all five use cases, enabling real-time decision-making.

- **Use Case 1: Crop Yield Prediction**

Using FAOSTAT and rainfall datasets, a linear regression model predicted yield with an R^2 score of 0.89 and RMSE of 120.5. The relationship between rainfall and yield was strong, showing an optimal rainfall range of 1100–1400 mm for higher productivity.

- **Use Case 2: Rainfall Impact Analysis**

Rainfall variations were strongly correlated with yield (r = 0.83). Years with below-average rainfall showed a notable yield decline. Adaptive irrigation systems are essential to mitigate rainfall fluctuations.

- **Use Case 3: Crop-Specific Climate Matching**

K-Means clustering identified crop–climate zones:

- Rice: 25–32°C, >1200 mm rainfall
- Wheat: 15–25°C, 800–1000 mm rainfall
- Maize: 20–28°C, 700–1200 mm rainfall
  These insights help policymakers plan region-specific crop allocation.

- **Use Case 4: Growing Degree Days (GDD)**

The model estimated required GDD for crop maturity—Rice: 1600–1800, Wheat: 1400–1600. GDD assists farmers in optimizing planting and harvesting periods to maintain consistency in yield.

- **Use Case 5: Soil Type Recommendation**

A Random Forest model achieved 93% accuracy in predicting suitable crops for given soil types. Nitrogen, pH, and organic carbon were the most influential parameters. Example outputs:

- Loamy Soil → Rice, Maize
- Sandy Soil → Groundnut, Millet
- Clay Soil → Paddy, Sugarcane

- **Visualization Insights**

The Databricks dashboard displayed yield trends, rainfall-yield correlations, and soil–crop heatmaps, supporting actionable, data-driven agricultural planning.

Chapter 5: Conclusion and Future Scope

- **Conclusion**

The Agri Data Scope project effectively applied Big Data Analytics and machine learning to the agricultural domain. Through Databricks, PySpark, and MLlib, a robust data pipeline was designed to integrate multiple datasets and generate predictive insights. The study demonstrated that Big Data frameworks can revolutionize farming practices through scalable data integration and predictive modelling.

- **Key Achievements**

1. Developed a multi-layer Big Data pipeline (Bronze–Silver–Feature–Model–Visualization).
2. Implemented five analytical models for key agricultural use cases.
3. Built a Databricks dashboard for visualization and policy decision support.
4. Enhanced model accuracy and data reliability using Delta Lake and MLlib tools.

- **Challenges**

1. Data inconsistencies across sources
2. Missing values and limited regional data coverage
3. Parameter tuning complexity in PySpark ML models
4. Limited computational resources during training

- **Future Scope**

1. Integrate IoT-based real-time data (temperature, humidity, soil moisture).
2. Incorporate geospatial and satellite imagery for crop health assessment.
3. Apply advanced ML algorithms (LSTM, ARIMA) for time-series forecasting.
4. Deploy as a cloud-based agricultural decision support system for broader accessibility.
5. Implement automated farmer alerts for rainfall, irrigation, and pest warnings.

- **Summary**

Big Data Analytics offers immense potential for advancing sustainable agriculture. The Databricks-based system developed in this project demonstrates how agricultural data can be transformed into meaningful intelligence to enhance yield prediction, climate adaptation, and resource optimization. This framework can guide future smart farming initiatives and policy formulations aimed at ensuring food security and environmental sustainability.