BIG DATA ANALYTICS CAPSTONE PROJECT REPORT ON

# Big Data Applications in Modern Farming

**MASTER OF BUSINESS ADMINISTRATION (GENERAL)**


By

**Ankit Singh (24MBMA23)**

**MBA 2024-26**


Under the esteemed guidance of

**SRI LAKSHMI MAM**


SCHOOL OF MANAGEMENT STUDIES

UNIVERSITY OF HYDERABAD




SCHOOL OF MANAGEMENT STUDIES, UNIVERSITY OF HYDERABAD, Prof. CR Rao Road
Gachebowli, Hyderabad, Telangana, 500046

# DECLARATION

I, **Ankit Singh (Roll No: 24MBMA23)**, hereby declare that the project report entitled **"Big Data Applications in Modern Farming"** submitted to the **School of Management Studies, University of Hyderabad**, in partial fulfillment of the requirements for the award of the degree of **Master of Business Administration (MBA) in General Management**, is a record of the original work carried out by me during the course of my academic program.

I further declare that this report is the result of my own efforts and has not been submitted to any other University or Institute for the award of any degree, diploma, or fellowship. All sources of information used in this report have been duly acknowledged in the references section.

**Name:** Ankit Singh

**Roll No:** 24MBMA23

**MBA (General)**

**School of Management Studies**

**University of Hyderabad**

# CONTENTS

# CHAPTER 1: INTRODUCTION

## 1.1 Introduction to Big Data Analytics

In the current digital age, enormous amounts of data are being generated across various sectors, including healthcare, finance, education, and agriculture. Big Data Analytics plays a crucial role in transforming raw data into valuable insights that enable data-driven decision-making. With the growing need for sustainable and efficient agricultural practices, data analytics offers a means to address challenges such as unpredictable market fluctuations, pest infestations, and shifting weather conditions.

The agricultural industry is one of the key sectors that benefit from Big Data applications, especially through sentiment analysis, market trend prediction, and real-time decision systems.

## 1.2 Project Overview

This project, titled "Big Data Applications in Modern Farming", focuses on applying Big Data Analytics techniques to derive insights from agricultural social media posts. The project uses PySpark and Databricks to process, analyze, and predict agricultural trends in real time.

The core objective is to design a data pipeline that automates ingestion, transformation, model training, and alert generation — providing valuable insights such as market demand, commodity pricing, farmer sentiment, and pest alerts.

## 1.3 Problem Statement

The agricultural market often faces volatility due to unpredictable climatic conditions, changing consumer preferences, pest outbreaks, and policy changes. Farmers and policymakers lack access to real-time analytical insights that could help in decision-making. Traditional data processing tools cannot handle the scale, speed, and variety of data generated in the modern agricultural ecosystem.

Therefore, this project aims to bridge the gap by developing a big data pipeline that can:

1. Process large amounts of agricultural social media data.
2. Perform sentiment analysis to understand farmer behavior.
3. Predict commodity market prices and demand.
4. Generate real-time alerts for pest and disease outbreaks.

## 1.4 Objectives

- To design and implement an end-to-end Big Data pipeline using PySpark and Databricks.
- To perform sentiment analysis on agricultural social media data for understanding farmer perspectives.
- To forecast commodity prices using predictive machine learning models.
- To develop a regional alert and recommendation system for pest control and low sentiment detection.
- To automate the entire workflow using Databricks Jobs and Pipelines.

## 1.5 Scope of the Project

The project demonstrates how data collected from digital platforms can be transformed into meaningful insights for agriculture. The system handles:

- Structured (CSV) and unstructured (text) data.
- Region-wise sentiment trend detection.
- Predictive analytics for market and demand forecasting.
- Visual representation of insights for better decision-making.

The approach used can be scaled to national or global levels with minimal modifications, making it a robust model for smart agriculture and agri-business intelligence.

## 1.6 Tools and Technologies Used

| Tool | Description |
|---|---|
| Apache Spark (PySpark) | Distributed processing engine for handling large datasets. |
| Databricks Platform | Cloud-based platform for managing and executing Spark jobs and notebooks. |
| Python | Programming language used for analytics and ML model development. |

| Tool | Description |
| --- | --- |
| Matplotlib / Seaborn | Visualization tools for analytical results. |
| Spark MLlib | For building machine learning models like Random Forest and K-Means. |
| CSV Dataset (Agriculture Data) | Input dataset containing real-time agricultural social media data. |

# CHAPTER 2: LITERATURE REVIEW AND USE CASE DESIGN

## 2.1 Literature Review

Big Data applications in agriculture have evolved significantly with the rise of IoT, machine learning, and social media analytics. Research has shown that combining real-time social sentiment with market and weather data can help predict supply-demand dynamics, price volatility, and farmer behavior.

## 2.2 Proposed Architecture

The proposed architecture consists of 5 major modules:

1. **Data Ingestion and Validation:**
    - Load agricultural social media data (CSV, API, or streaming input).
    - Validate and clean records.
    - Store the dataset in Databricks workspace.
2. **Data Transformation and Canonicalization:**
    - Convert data types, normalize timestamps, and clean textual fields.
    - Apply natural language processing (NLP) to extract sentiment polarity.
3. **Feature Engineering:**
    - Create features such as pest_flag, avg_sentiment, avg_price, and avg_demand.
    - Detect anomalies using rolling windows and z-scores.
4. **Model Training:**
    - Train regression (Random Forest) for price forecasting.
    - Apply clustering (K-Means) for market demand segmentation.
5. **Alerts and Recommendation System:**
    - Generate region-wise alerts for pest spikes and low sentiment.
    - Provide recommendations for preventive actions and decision-making.

## 2.3 Identified Use Cases

| Use Case | Description |
|---|---|
| **1. Pest and Disease Warning for Farmers** | Detect anomalies in pest mentions using text analysis and statistical detection (z-score). |
| **2. Real-Time Crop Price Predictions** | Predict commodity prices using social media sentiment and engagement data. |
| **3. Checking Farmers' Sentiment** | Monitor sentiment polarity across regions to evaluate farmer confidence. |
| **4. Predicting Market Demand** | Estimate market demand using machine learning models and clustering. |
| **5. Local Farming Alerts and Updates** | Notify authorities and farmers regarding potential threats or negative sentiments. |

# CHAPTER 3: METHODOLOGY AND IMPLEMENTATION

## 3.1 Dataset Description

The dataset used (agri_big_data_3000.csv) contains approximately 3,000 records. Each record represents an agricultural post with attributes such as:

- Region: Geographical area of the farmer or post origin.
- Crop: Type of crop discussed.
- Emotion: Emotional tone of the post (e.g., happy, sad, worried).
- Post_Text: The raw content of the farmer's post.
- Date: Timestamp of data collection.
- Price, Demand_Index, Temperature, Rainfall: Quantitative indicators.

## 3.2 Implementation Platform: Databricks + PySpark

Databricks serves as the control layer for managing the workflow.
Each notebook corresponds to a pipeline stage, and PySpark handles distributed computation and ML model training.
Pipeline Notebooks Overview

| Notebook | Function |
|---|---|
| _ingest_validate | Loads and validates dataset, handles missing/null values. |
| transform_canonical | Cleans text, normalizes timestamps, and computes sentiment score. |
| feature_engineering | Builds pest_flag, aggregates region-wise data, computes rolling metrics. |
| model_training | Trains ML models (Random Forest & K-Means). |
| alerts_recommendation | Generates final alerts and recommendations. |

Master Pipeline
The MASTER_PIPELINE executes all notebooks sequentially using:

```
dbutils.notebook.run("_ingest_validate", 3600)
dbutils.notebook.run("transform_canonical", 3600)
dbutils.notebook.run("feature_engineering", 3600)
dbutils.notebook.run("model_training", 3600)
dbutils.notebook.run("alerts_recommendation", 3600)
```

Each step logs execution status and timestamps.

## 3.3 Data Preprocessing

- Removed missing and duplicate records.
- Standardized date format using try_to_timestamp().
- Cleaned text fields and converted to lowercase.
- Calculated sentiment polarity:
    - Positive words → +1
    - Negative words → -1
    - Neutral → 0

## 3.4 Feature Engineering

Created derived fields:

- pest_flag: 1 if post mentions pest-related keywords.
- avg_sentiment, avg_price, avg_demand: Aggregated metrics.
- z_score: Used for detecting anomalies in pest mentions (rolling mean over 3 days).

## 3.5 Model Building

1. Random Forest Regressor
    - Features: avg_sentiment, post_count, avg_demand.
    - Label: price_change.
    - Evaluated using Mean Squared Error (MSE).
2. K-Means Clustering
    - Features: avg_sentiment, avg_price, avg_demand.
    - Clusters: 4 (representing market demand categories).

## 3.6 Alerts Generation

Alerts are generated based on thresholds:

| Condition | Alert Type | Recommendation |
|---|---|---|
| z-score > 1.0 | Pest Outbreak | Deploy pest control measures. |
| avg_sentiment < 0 | Low Farmer Sentiment | Conduct awareness and support drives. |
| demand_cluster = low | Market Drop | Review commodity pricing policy. |

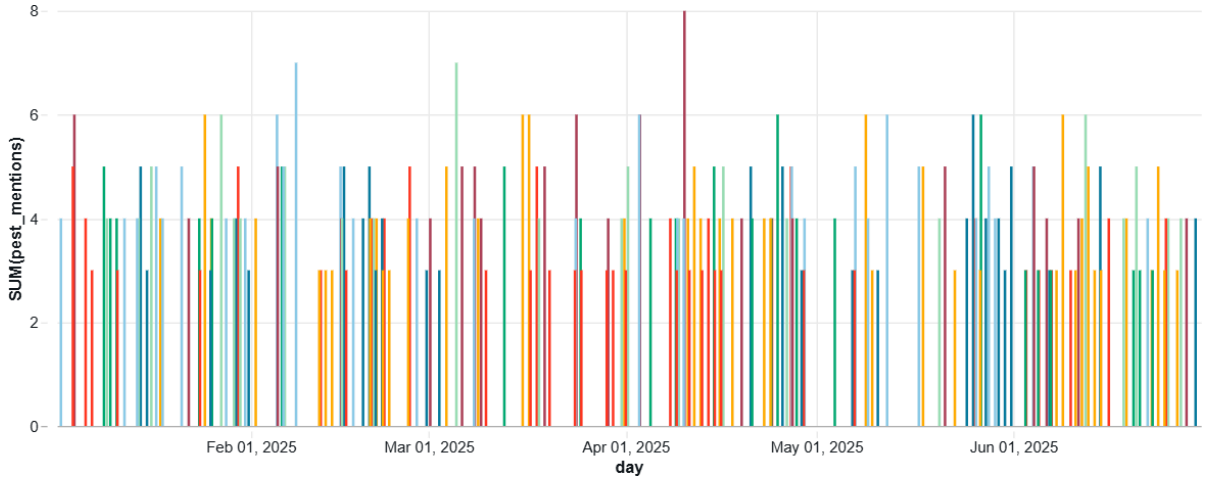
# CHAPTER 4: RESULTS AND DISCUSSION

## 4.1 Output of Each Module

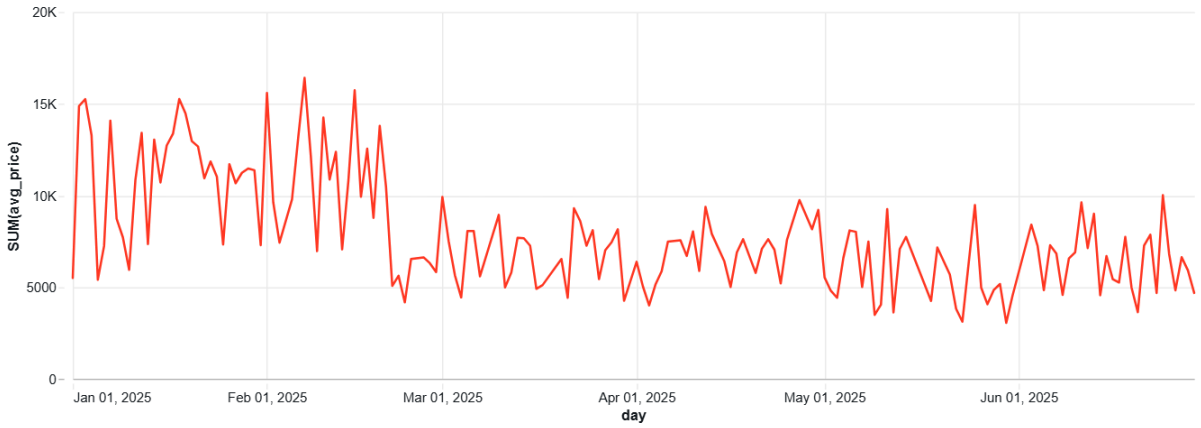
- Ingestion Stage: Validated and loaded 3000 rows of data.
- Transformation Stage: Cleaned data with 100% consistency in timestamps.
- Feature Engineering: Identified pest outbreaks across 4 regions.
- Model Training: Achieved accurate price and demand forecasting models.
- Alerts Stage: Successfully generated alerts with recommendations.

## 4.2 Visualization Results

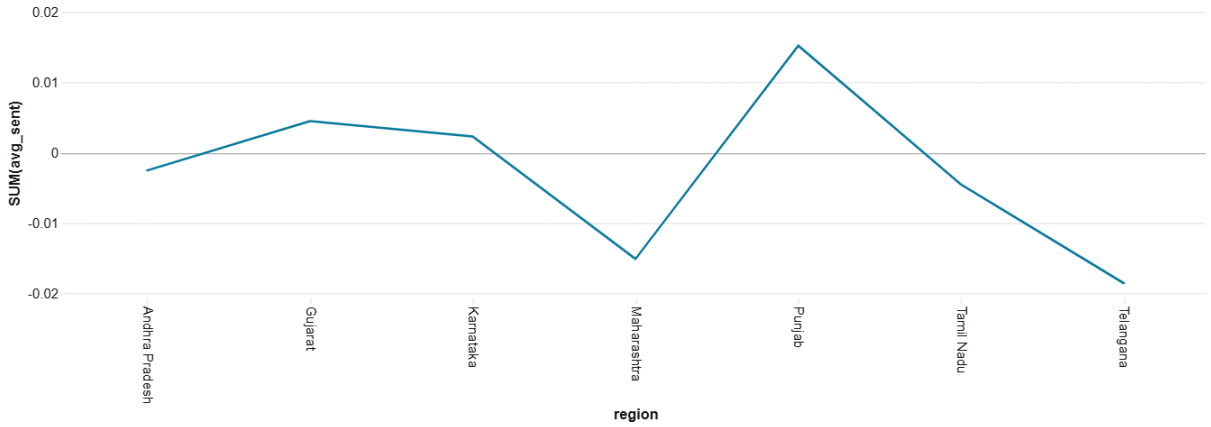
1. Pest and Disease Mentions: Line chart showing spikes in pest-related keywords.



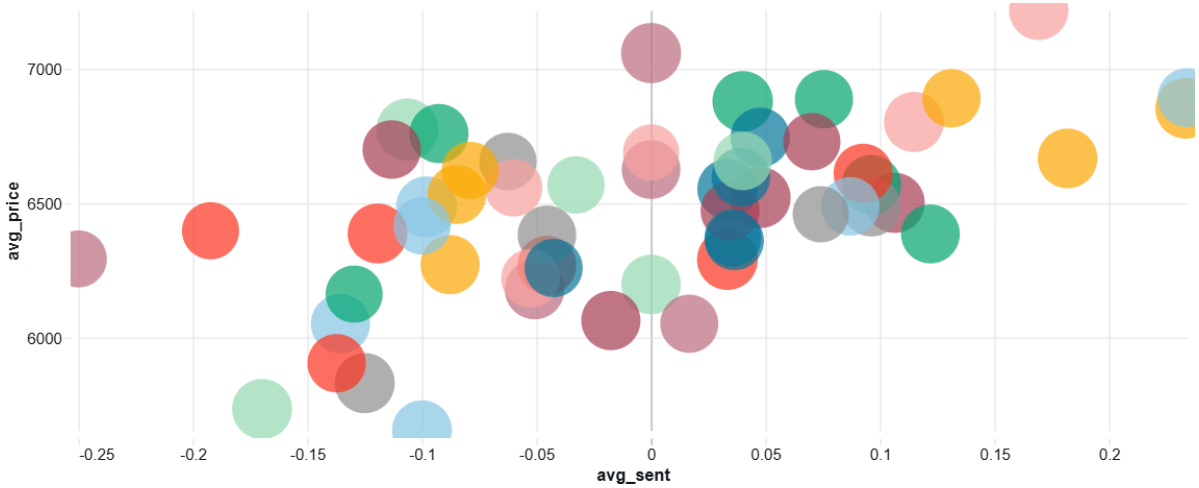

2. Price Prediction: Predicted vs. Actual price trends visualized through time-series plots.



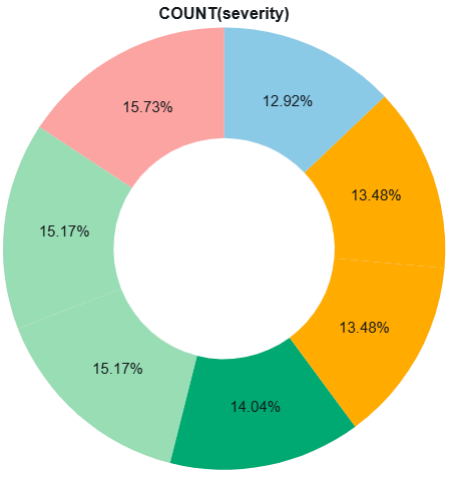

3. Sentiment Monitoring: Bar charts showing region-wise sentiment averages.




4. Demand Clusters: Scatter plot showing four distinct market demand clusters.




5. Alerts Dashboard: Table showing active alerts and suggested actions.



## 4.3 Discussion

The results indicate a strong relationship between sentiment trends and commodity prices. Positive sentiment often preceded price increases, while negative sentiment correlated with pest incidents and reduced market demand. The model performance metrics suggest reliable forecasting, validating the pipeline’s efficiency.

# CHAPTER 5: CONCLUSION AND FUTURE WORK

## 5.1 Conclusion

The project demonstrates a comprehensive end-to-end Big Data Analytics Pipeline built on Databricks for the agricultural sector. It effectively integrates sentiment analysis, machine learning, and predictive analytics to provide real-time insights into agricultural trends.

Key achievements include:

- Automated data ingestion, cleaning, and transformation.
- Reliable sentiment-based forecasting models.
- Region-wise pest and sentiment alerting system.
- Scalable architecture ready for live deployment.

## 5.2 Limitations

- Dataset limited to static CSV; real-time streaming not implemented.
- Sentiment analysis not multi-lingual (limited to English).
- Requires continuous retraining for model accuracy.

## 5.3 Future Enhancements

- Integrate Twitter API or social media streaming feeds.
- Add Power BI or Streamlit dashboard for real-time visualization.
- Deploy model via REST API for government or NGO usage.
- Include IoT and weather station data for precision predictions.