Final Project Report

Internship Program – upSkill Campus & Uniconverge Technologies (P) Ltd.

Project Title: Prediction of Agriculture Crop Production in India

Submitted by: M.Mohammed Salman

B.Tech (AI & DS), B.S. Abdur Rahman Crescent Institute of Science and Technology

Submission Date: 31 July 2025

Table of Contents

1. Introduction

Agriculture forms the backbone of India's economy, with a majority of its population dependent on farming activities.
Accurate crop production prediction plays a vital role in policymaking, resource allocation, and farmer decision-making.
This internship project aimed to apply Machine Learning techniques to forecast agricultural crop production in India using historical datasets.

2. Problem Statement

The agricultural sector faces multiple challenges including unpredictable weather, pest attacks, and resource mismanagement.
These challenges lead to fluctuating crop yields. The project's objective is to develop a machine learning-based prediction system
that can forecast crop production trends and assist in improving efficiency and planning.

3. Dataset Description

The dataset was sourced from data.gov.in, covering the period between 2001 and 2014. It includes information about crop names,
varieties, states, quantities, production levels, seasons, cost of cultivation, and recommended zones.
This rich dataset provides opportunities to analyze crop performance across states and over time.

Key columns include: Crop, Variety, State, Quantity, Production, Season, Unit, Cost, and Recommended Zone.

4. Methodology

The project followed a structured workflow:
- Data Preprocessing: Cleaning missing values, handling inconsistent formats, and encoding categorical variables.
- Exploratory Data Analysis (EDA): Understanding crop patterns by state, season, and variety using graphs and heatmaps.
- Feature Engineering: Derived new features like cost-per-ton and production-per-hectare

to improve prediction accuracy.
- Model Building: Implemented Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor models.
- Model Evaluation: Performance was evaluated using metrics like R^2 score and RMSE. Random Forest and Gradient Boosting showed the best performance.

5. Results & Discussion

The analysis revealed distinct crop patterns across states and seasons. Wheat and rice production dominated across major states like Uttar Pradesh and Punjab.
Model performance results were as follows:
- Linear Regression: $R^2 = 0.72$
- Random Forest Regressor: $R^2 = 0.87$
- Gradient Boosting Regressor: $R^2 = 0.85$

Random Forest was selected as the most reliable model. SHAP values were applied to explain feature importance,
which highlighted cost and state-wise cultivation area as key predictors.

6. Challenges Faced

- Data imbalance across states led to skewed predictions for minority crops.
- Overfitting occurred in early models, requiring cross-validation and tuning.
- Feature importance varied across models, creating interpretation challenges.

7. Lessons Learned

- Practical knowledge of preprocessing and handling messy datasets.
- Experience with hyperparameter tuning to improve model performance.
- Use of SHAP for model interpretability to identify significant features.
- Importance of structured reporting and project documentation for professional deliverables.

8. Conclusion & Future Scope

The project successfully developed a machine learning pipeline to predict crop production in India with high accuracy.

Future work could involve integrating weather and soil quality datasets, as well as developing real-time dashboards for stakeholders.
This approach could provide actionable insights for policymakers and farmers to optimize resource allocation and improve yield outcomes.

9. References

- Dataset: https://data.gov.in/
- Tutorials on Regression Models and SHAP values
- Documentation on Random Forest and Gradient Boosting methods