Why Do Some Districts in Kerala Produce More Than Others?

Project Report For Data Science

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Under the Guidance of

DR. PIYUSH CHAUHAN

Associate Professor

Department of Computer Science & Engineering

Symbiosis Institute of Technology,

Nagpur Campus

Course Name: Data Science

Submitted by

Sujal S Acharya

VII SEM

PRN: 22070521037

B. Tech Computer Science & Engineering

Symbiosis Institute of Technology,

Nagpur Campus

Table of Contents

1. Project Summary

Agriculture plays a central role in Kerala's economic and social development. However, despite shared geographical and cultural similarities, different districts in the state exhibit significant variation in crop productivity, agricultural diversity, and resource utilization. This project investigates these variations and aims to answer the question: **Why do some districts in Kerala produce more than others?** The study integrates exploratory data analysis (EDA), clustering analysis, and machine learning-based yield prediction to identify the factors influencing agricultural disparity. By combining insights from multiple analytical methods, the project contributes to strategic agricultural decision-making and supports data-driven policy frameworks.

Key Objectives

- Understand crop area, production, and yield variations across districts.
- Identify natural groupings of districts based on agricultural performance.
- Predict district-crop yield categories (Low/Medium/High) using machine learning.
- Support agricultural planning and policymaking through analytical evidence.

2. Data Sources

The project uses consolidated agricultural datasets containing district-wise values for crop production, cultivated area, yield, crop types, soil categories, seasons, and related environmental variables. The datasets were merged and cleaned from multiple notebook sources to create a uniform analytical foundation.

Data Components

- District-wise cultivated land area (hectares)
- Crop production quantity (metric tons)
- Yield values (kg/ha)
- Crop category and type
- Season and soil classification
- Encoded ML-ready dataset

Preprocessing Performed

- Handling missing and null values
- Standardizing inconsistent text entries
- Encoding categorical fields (LabelEncoder, One-Hot)
- Removing duplicate and anomalous rows
- Normalizing numerical values (StandardScaler)

3. Project Setup

The project was conducted in Google Colab / Jupyter Notebook using Python. The environment allows efficient data handling, real-time visualization, and interactive model experimentation.

Tools & Libraries Used

- Data Handling: Pandas, NumPy
- Visualization: Matplotlib, Seaborn, Plotly Express
- Machine Learning: Scikit-Learn (DecisionTreeClassifier, KMeans, PCA, encoders, scalers)
- Development Environment: Colab/Jupyter Notebook

Reasons for Tool Selection

- Flexibility for iterative testing
- Strong visualization support
- Reproducible environment with structured workflow
- Easy integration into dashboards or production systems

4. Scope & Methodology

This project follows a structured multi-phase analytical framework designed to extract insights and build predictive capability.

Workflow Stages

Data Cleaning

- Dealing with null values and outliers
- Standardizing names and numeric formats
- Encoding and scaling features

Exploratory Data Analysis

- Trend visualization and comparative charts
- Correlation heatmaps for relationship discovery

Feature Engineering

- Derived metrics such as total production and crop diversity
- Generation of ML-ready feature sets

Clustering

- K-Means clustering to identify agricultural groups
- PCA for dimensionality reduction and visualization

Machine Learning Classification

- Decision Tree model to predict yield category
- Accuracy, confusion matrix, and classification report evaluation

5. Exploratory Data Analysis

EDA helped uncover major patterns and variability in production efficiency among districts. Visual analysis demonstrated significant differences in crop selection, seasonal output, and yield stability.

Key Observations

- Some districts consistently show higher production due to better land availability and resource utilization.
- Crop diversity strongly correlates with production stability.
- Positive relationship observed between cultivated area and production, but anomalies exist due to climatic variability and resource constraints.
- Certain crops produce high yield even with small land allocation.

Types of Visualization Used

- Bar charts comparing district-wise production
- Area vs production scatterplots
- Heatmaps showing feature correlation strength
- Line charts displaying crop and yield trends

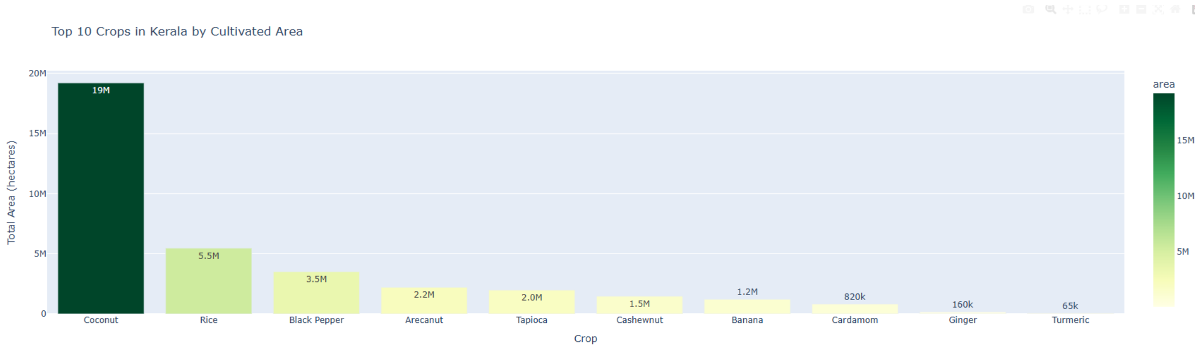


Fig 1 : Top 10 cultivated Crops in Kerala

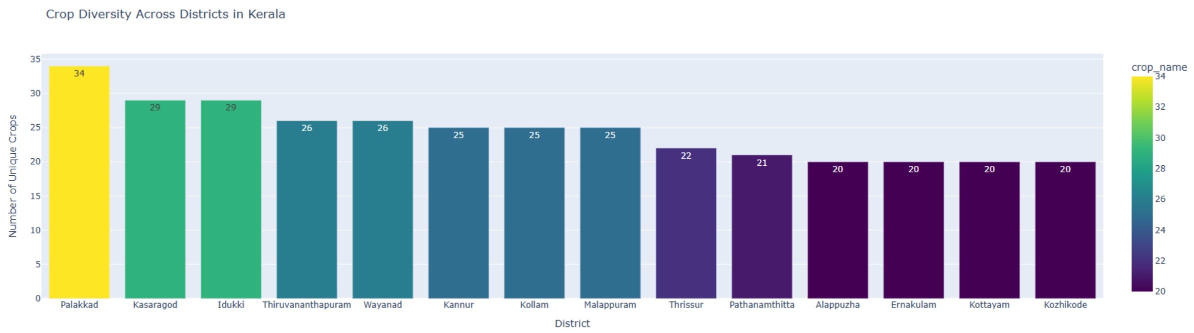


Fig 2 : Crop Diversity Across Kerala

6. District Clustering

Clustering using K-Means grouped Kerala districts into three meaningful performance-based clusters, enabling more targeted agricultural intervention.

Cluster Characteristics

Cluster	Description	District Type
Cluster 1	High production, large cultivated area, high crop diversity	Dominant agricultural districts
Cluster 2	Moderate production, balanced crop mix, average land use	Stable developing districts
Cluster 3	Low production, limited area, low diversification	Resource-limited districts

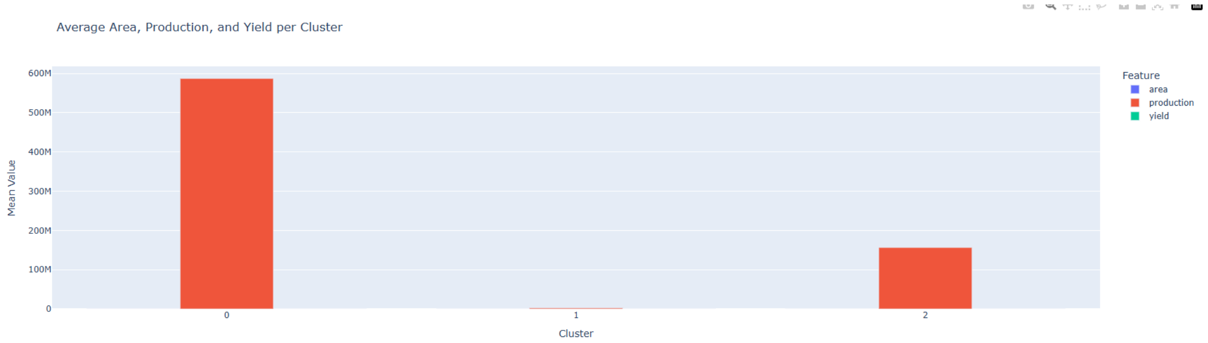


Fig 3 : Clusters information

7. Machine Learning Models

A Decision Tree Classifier was trained to predict yield level categories: Low, Medium, and High. The model analyzed crop-environment relationships and evaluated feature contributions.

Model Pipeline

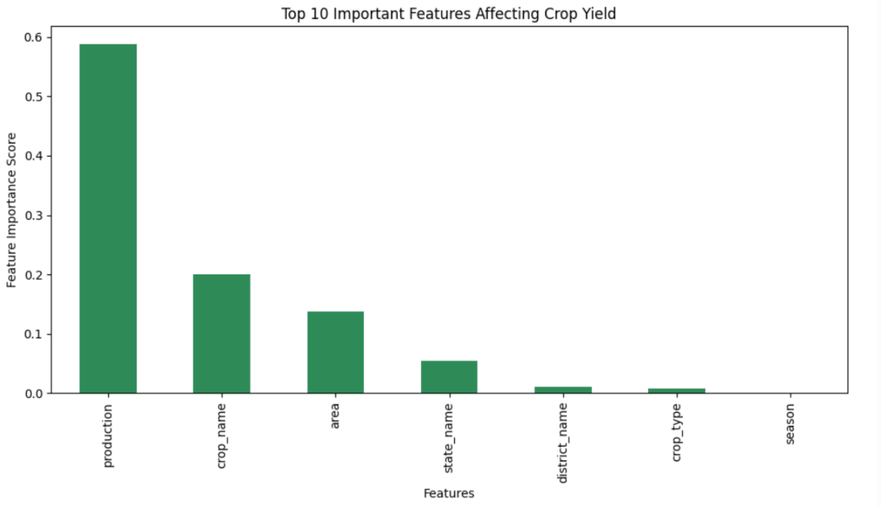
Label Encoding → Scaling → Train-Test Split → Model Training → Evaluation

Reasons for Model Selection

- Simple interpretability and decision rule visualization
- Handles categorical and continuous features well
- Performs effectively on tabular datasets

Major Feature Contributors

- Cultivated Area
- Total Production
- Crop Type
- Soil Type
- Season



8. Key Combined Insights

Strategic Findings

- Land area alone does not determine productivity; optimized crop selection matters.
- Crop diversity increases resilience and output stability.
- Significant agricultural inequality exists between districts.
- Machine learning can successfully predict yield behavior.
- Region-specific planning is essential to maximize growth.

Applications

- Government planning and subsidy allocation
- Farmer decision-support systems
- Early warning yield forecasting tools

9. Conclusion

This study presents an end-to-end analytical framework for understanding agricultural variation across Kerala. Through integrated EDA, clustering, and machine learning, it identifies performance drivers and predicts future outcomes. The results underline the importance of diversification, scientific farming strategies, and data-supported policy decisions. The project lays the groundwork for expansion into forecasting models and real-time agricultural dashboards that could enhance decision-making efficiency and support sustainable crop practices.