SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR
Constituent of Symbiosis International (Deemed University), Pune
(Established under Section 3 of the UGC Act of 1956 wide notification number F-9-12/2001-U-3 of Government of India)
॥वसुधैव कुटुम्बकम्॥ Re-Accredited by NAAC with 'A++' Grade

Cost of Cultivation Analysis and Crop Yield Classification

Exploratory Data Analysis and Trend Insights – PMKSY Progress Monitoring Dataset: An Analytical Study of Financial and Physical Targets vs Achievements Across Indian States and Time Periods

MINI PROJECT REPORT

Submitted To:

Dr. Piyush Chauhan

Submitted By:

Sakshi Parate

22070521074

VIIth sem, C

1. TABLE OF CONTENTS

2. ABSTRACT

The selected project is the exploratory data analysis (EDA) and predictive modeling of the Cost of Cultivation data collection found at the Government of India Agricultural Statistics Division (Data.gov.in). The data has comprehensive data on agricultural inputs like fertilizers, seeds, human and animal labor, and irrigation expenditures on different crops, states and years. The purpose of the research is to examine the relationship between various cost parameters and crop yield and group yield levels as Low, Medium, and High with the use of Machine Learning and Convolutional Neural Network (CNN) models. The analysis was done using univariate and bivariate exploration to see how the features are distributed and how the variables relate to each other, and correlation heatmaps were used to see which features have significant relationships. Models such as Linear Regression, Ridge, Lasso, KNN, and Random Forest were put into place after preprocessing steps such as data cleaning, encoding, and feature scaling with the best regression performance obtained by the Random Forest, and the highest classification accuracy obtained by CNN. This paper brings out the influence of different cultivation costs on yield effects and offers an insight that can be utilized by policymakers and farmers to manage agricultural investments, cost effectiveness, and sustainable crop productivity in India.

3. KEYWORDS

Cost of Cultivation, Data Analytics, Machine Learning, Predictive Modeling, Data Visualization, Agricultural Data Mining, Crop Yield Classification, Convolutional Neural Network (CNN), Regression Analysis, Univariate and Bivariate Analysis, Agricultural Data Analysis, Yield Prediction, Feature Correlation

4. INTRODUCTION

The Indian economy is based on agriculture as it forms a great part of the employment and national income source. Farm Management Cultivating Cost is also a critical factor in achieving profitability of farms, policy decision-making, and the sustainability of agriculture in general. The Government of India, via the different department of agriculture, avails comprehensive statistics on inputs cost and yield performance of various crops in different states and time. This information is an important asset in learning how the cost of production, the use of inputs, and the geographical differences influence the yield of agriculture.

We conducted the Exploratory Data Analysis (EDA) in this project on the publicly available dataset on the Cost of Cultivation that was on Government of India (Data.gov.in). The analysis discusses the cost elements like labour, fertilizer, seed, irrigation and use of machinery and how they are related to derived crop yield. We have tried to predict yield results as well as classify crop yields with the help of the models of Machine Learning (ML) and Deep Learning (CNN), in the categories of Low, Medium and High. This assists in establishing the important variables that can affect productivity, as well as building factual information to enhance farm productivity and decision-making.

4.1 Why I chose this dataset

The reason why the Cost of Cultivation data set was selected is because it is an agricultural and economic country of interest that provides a comprehensive picture of the expenses of crop production and yield performance in India. It is maintained officially by the Ministry of Agriculture and Farmers Welfare, which secures credibility and in-depth in both cost structure and data of productivity. This data set gives an option of analyzing and predicting efficiency of agriculture, optimization of costs and yield improvement.

The reasons why this data set has been selected are as follows:

- **National Importance: Gives insight into the cost effectiveness, profitability and yield difference across states and crops in India**.
- **Authority: Real information through government checked data with standardized regularly updated data**.
- Rich in Features: Has detailed input costs such as fertilizer, seed, labor, irrigation and derived yield.
- Machine Learning Scope: Allows regression, classification and CNN-based predictive modeling.
- Practical Value: Assists in finding high-cost and low-yield areas which helps farmers and policymakers to make improved resource allocation.

4.2 Dataset Description

The data utilized in this project is obtained in Government of India - Data.gov.in at the Cost of Cultivation statistics in the Ministry of Agriculture and Farmers Welfare. This is a data set that has a variety of information on the cost structure and the productivity of all the crops grown in the different Indian states. It includes economic and physical indicators that are useful in determining the impact of input costs on yield outcomes and the profitability of a farm as a whole.

- Source: Ministry of Agriculture & Farmers Welfare, Government of India (Data.gov.in)
- Coverage: Indian states and crops major.
- Time Period: The period is about 2000-2020 financial years.
- Granularity: State and crop cost and yield information.
- Format: Tabular information that is structured and has both categorical and numerical variables.

4.3 What the Dataset Contains

The Cost of Cultivation data set covers all the economical and physical features of agricultural crops in the different states in India. It dwells on the finer cost elements that are entailed in the production of various crops and the output. The dataset is a composite of various cost types- inputs, labor and operation costs and productivity measures. This information assists in the analysis of the correlation between the cost of inputs, the farming processes, and the agricultural production.

- Time-based data: Encompasses several financial years (around 2000-2020).
- Geographical data: Names and codes of Indian states.
- Cost data: Details: information about different costs of cultivation and production (e.g., A1, A2, B1, B2, C1, C2, C3).
- Output information: Value of the main product, value of the by-product and derived yield.
- Analytical value: The value may be used to make comparisons of state costs efficiency, profitability, and productivity as well as crop comparisons.

Table 1. Key Columns in the Dataset

Sr.no	Column Name	Description
1.	id	Unique identifier for each record
2.	year	Financial or calendar year of record
3.	state_name	Name of the state
4.	state_code	State code abbreviation
5.	crop_name	Name of the crop

6.	crop_code	Unique crop identifier
7.	crop_type	Type/ category of crop (e.g., cereal, pulse, oilseed)
8.	cul_cost_a1–c2rev	Various cultivation cost categories
9.	prod_cost_a1–c3	Production cost categories
10.	main_prod_value	Value of the main agricultural product
11.	by_prod_value	Value earned from crop by-products
12.	mat_lab_input_seed/fertilizer/labor	Input material and labor costs
13.	Derived_yeild	Calculated yield per hectare or unit
14.	yield_class	Target variable indicating yield category (Low, Medium, High)

4.4 Problem Statement

Despite the detailed information about the costs involved in cultivation and the yields, the agricultural cost datasets are not used to the fullest as the systematic analytical modelling is missing. Lacking an in-depth analysis of data, it is hard to know which elements of cost have the greatest impact on crop productivity or how the distribution of resources impacts yield outputs. Thus, the most crucial issue that the given project will solve is the lack of predictive analytical framework, which has the potential to integrate cultivation costs and yield performance and classify the level of productivity based on the methods of machine learning.

4.5 Project Objectives

The key objectives of this project are as follows:

- To conduct an exploratory data analysis (EDA) on the Cost of Cultivation data to comprehend data organization and other important agricultural trends.
- To utilize machine learning models on yield prediction and classification including Linear Regression, Ridge, Lasso, KNN, Random Forest, and CNN.
- To determine the association between the cost of cultivation (labor and fertilizer, seed, irrigation) and yield values obtained. To compare and contrast the performance of models they need to be evaluated using such statistical measures as R2, RMSE, and accuracy.
- To come up with insights that are meaningful that can help farmers, policymakers and researchers to optimize resource use and better results in terms of productivity.

4.6 Novelty of the Research

The paper is unique since it combines the conventional machine learning methods and deep learning (CNN) to predict yield by using farm costs. In contrast to the earlier descriptive analyses, the present project is a combination of a regression-based and classification-based modeling with visual insights that would determine the most significant factors influencing crop yield. Its value is in the combination of economic data (cost of cultivation) and predictive modeling, which provides an opportunity to make agricultural decisions on the basis of data, forecast resources in the future and determine policies in advance.

5. LITERATURE REVIEW/RELATED WORK

Table 2. Empirical review of existing methods

Reference	Method Used	Findings	Results	Limitions
Yethiraj N. G. (2012)	A review on agricultural data mining.	Shows the use of analytics in enhancing farming decisions	There are noted advantages in predicting and monitoring	Conceptual; no implementation of data set
Ramesh D. & Vardhan B. (2015)	K-Means and Regression to crop yield analysis	The type of soil and rainfall that is found influence crop performance	Suited to the analysis of yield variations	Inaccurate data on crops; precision due to lack of data.
Manjula E. & Djodiltachoumy S. (2017)	Crop yield prediction model Regression based.	Past information enhances accuracy of predictions.	Grew dependable outputs of large crops.	Localized; viability untested.
Dhivya B. et al. (2017)	ML and data mining survey.	Comparative analysis of yield prediction algorithms.	Assisted in the choice of appropriate predictive techniques.	Absent real world experimental validation.
Majumdar J. et al. (2017)	Agricultural datasets Big data mining and clustering.	Determined clustering patterns of productivity.	Improved knowledge of yield changes	Large data computationally difficult.
L. R. Beldarrain et al. (2021)	The effect of ageing on meat quality statistically.	Found ageing influences the fatty acid structure and the meat quality.	Gave measureable information on the effects of ageing on meat quality.	Domain - specifi c; to irrigation data not applicable.
Nishant P. S. et al. (2020)	Crop yield prediction Machine Learning models.	The results of yield depend on climate and soil factors.	High accuracy on yield forecasting.	Small size of dataset and variables.
Jain R., Kishore P., & Singh D. (2020)	Irrigation data analysis based on statistics and policies.	Poor distribution of water and identified inefficiency.	Recommended combined water management plans.	None of the predictive or ML modeling used.

Patil S. & Kumar M. (2021)	Ridge and random forest cost based yield prediction	The predictors found include fertilizer, seed and labor cost.	Obtained an accuracy of about 95 in yield forecasting.	Localized data; not generalized on a national basis.
Beldarrain L. R. et al. (2021)	Quality of agricultural products, statistical analysis.	Yield is affected by found cost and input variables.	Gave measurable information on productivity.	Predictive not domain-specific.
Chethana Sridhar et al. (2022)	Optimized CNN that is used to imagine agriculture.	Enhanced crop and plant illnesses identification.	High precision and recall obtained	Only limited to data in images.
Sathya M. et al. (2022)	Genetic algorithm of screening agricultural features.	Chosen selected key yield-related factors.	Improved model performance using less variables.	Primarily used on biological data.
Pareek P. K. et al. (2022)	Smart image reduction optimization model.	Better accuracy and compression of image-based data.	Less computation time and information loss.	Concentrated on imaging; not tabular cost data.
Pai A. H. et al. (2022)	ML + IoT in real time crop monitoring.	Faster network performance and data.	Improved agriculture IoT systems throughput.	IoT; no yield prediction.
Schiano Di Cola et al. (2024)	Supervised learning EDA of plant phenotyping.	Strong correlations of characteristics and environment.	Better precision in classification of plant characteristics.	Small scope of data used; not experimented on big data.

Research Gap:

1. Absence of Exploratory Data Analysis of Cost of Cultivation Dataset:
    - Majority of the available literature centers on either predictive model or algorithmic optimization to forecast crop yield but does not pay much attention to the exploratory data analysis (EDA). The literature does not show any research carried out to determine the relationship between different factors of cultivation costs, which include fertilizer, labor, and irrigation, and the yield results of crops in different crop types and state in India.
2. Minimized Economic and Yield Data Integration:
    - Although some of the models forecast the yield on the basis of weather or soil data, not many of them incorporate economic variables such as cost of structure of input, value of by-products and operational costs. This loophole limits a comprehensive explanation

on the impact of financial and operational expenses on productivity and profitability.

3. Low adoption of Hybrid Deep Learning Methodologies:
    - There are many studies that have investigated the use of traditional models (Regression, KNN, Random Forest) to estimate yields, whereas limited studies have been conducted on the use of CNN-based models to analyze tabular agricultural cost data. Deep learning has not yet been fully utilized with regard to the potential to identify complex and nonlinear patterns in large-scale cultivation data.

4. Weakness in Regional and Temporal Analysis:
    - The majority of datasets or studies lack a comparison of yield patterns across time or region. An analysis of state-wise and crop-type based on cost features may assist in bringing about knowledge of differences in the regions, crop characteristics, and the trend of productivity with time.

6. METHODOLOGY / PROPOSED SYSTEM

The analysis of the dataset on Cost of Cultivation obtained by the Government of India was approached in a systematic way as an Exploratory Data Analysis (EDA), followed by a Machine Learning modeling. The data contains precise results about the cost of input as fertilizers, seeds, labor, irrigation and operational charges as well as resulted yield and by-products.

Following the data import, a cleaning and preprocessing of the dataset included the removal of missing data, feature normalization, and the encoding of categorical data to prepare it to be used in the model. The univariate and bivariate analyses were used in order to comprehend the distribution of cultivation costs and the relationship between them and yield outcomes. The research also applied various machine learning algorithms- Linear Regression, Ridge Regression, Lasso Regression, KNN Regression and Random Forest to predict yields, then a Convolutional Neural Network (CNN) to classify the yield (Low, Medium, or High). Meaningful insights and performance metrics, such as accuracy, precision, recall, and F1-score, were derived using visual tools, such as boxplots, correlation heatmaps, and trend plots.

6.1 Data Collection

- **Dataset Source:** Ministry of Agriculture and farmers Welfare, Government of India (Cost of cultivation Statistics).
- Link to dataset: https://indiadataportal.com/p/cost-of-cultivation/r/moafw-cost_of_cultivation-st-yr-dvq.
- The sample includes various financial years and seasons of crops in diverse states and crop varieties in India.
- Time Period: It includes some 2000-2020 financial years.
- It also contains specific data on the input expenses (fertilizers, seeds, labor, irrigation, and machinery), operations expenses, and obtained yield on each crop.
- This data was downloaded in CSV format and analyzed in Python with the help of Pandas.
- The simplest preprocessing steps were made to address missing values and standardization of the column names, encode categorical data (state name, crop name, crop type), and numerical uniformity across the cost variables.

6.2 Data Cleaning and Preprocessing

- Handling Missing Values:
  Eliminated rows that contained missing or unspecified values in both critical cost elements and yield-related columns to comply with the accuracy of the data.
- Equalization of Column Names and Formats:
  Assured standard naming conventions in the columns (i.e., culcostc2rev, prodcostc2rev, etc.) and categorical fields that are formatted (e.g. crop type, crop name, state name, and so on).
- Duplicate Removal:
  Identified and eliminated duplicate data to help prevent redundancy and achieve unbiased model training and analysis.

- Removing Invalid Entries:
  filtered the records with a zero or negative value of cost and unrealistic yield values that may misrepresent analysis and model fidelity.

- Basic Feature Creation:
  Developed derived features like yieldclass (Low, Medium, High) that were calculated on the derivedyield column to facilitate supervised classification and more analysis of yield performance.

6.3 Model Design / System Architecture

The entire mechanism will be aimed at converting the raw agricultural cost data into meaningful analytical information and predictive models of crop yield in the form of classification. The system has a systematic flow, which encompasses data ingestion, preprocessing, exploratory analysis, model training, and evaluation. All the steps play a role in providing accuracy, interpretability, and scalability of the findings.

- Data Source Layer: CSV data of the raw Cost of Cultivation data provided in the Ministry of Agriculture and Farmers Welfare, government of India.

- Processing Layer: The cleaning, encoding, normalization, and feature extraction of the data with the help of Python, Pandas, and NumPy to guarantee the data consistency and preparation to model.

- Exploratory Data Analysis (EDA): Univariate and bivariate analyses were performed with Seaborn and Matplotlib to comprehend the cost distributions, relations of features, and yields patterns across states and types of crops.

- Model Layer: Applied several Machine Learning models, Linear Regression, Ridge Regression, Lasso Regression, KNN Regression, and Random Forest Regression, as well as a Deep Learning model (CNN) in the classification of yield levels into Low, Medium, and High.

- Visualization Layer: Created plots like boxplot, bar graphs, correlation heat plots, and confusion matrices to feel the relationship, model accuracy, and key performance indicators.

6.4 Training and Evaluation

The data was split into training and testing data to both guarantee the model generalization and prevent overfitting. The models of Machine Learning and CNN were trained, validated, and evaluated with respect to performance metrics such as accuracy, precision, recall, and F1-score.

- Training Set (80%) - training and optimization of parameters of the model.

- Testing Set (20%) - applied to assess the model performance on unknown data.

6.5 Algorithms or Mathematical Formulations

1. Correlation analysis: It is used to determine the relationship between the key cost factors such as fertilizer, labor and seed costs and the yield obtained such as how many additional bushels per dollar of cost are obtained.

$$r = \frac{\Sigma\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{} \quad (1)$$

$$\sqrt{\Sigma\left(x_i-\bar{x}\right)^2\Sigma\left(y_i-\bar{y}\right)^2}$$

This assisted in determining the key elements of cost that affect agricultural yield enhancement the most.

2. Linear Regression Model: Regression model was applied to analyze the effect of different cost factors in yield prediction.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \quad (2)$$

where,

- Y = Predicted Crop Yeild
- $X_1, X_2, \ldots X_n$ = Independent features (e.g., Fertilizer Cost, Labor Cost, Seed Cost)
- β_i = Model coefficients (weights)
- β_0 = Intercept
- ϵ = Error term

Dependent and independent variables:

- **Independent Variables (X):** Cost-related factors (Fertilizer, Seed, Labor, Irrigation).
- **Dependent Variable (Y):** Derived yield (output of crop).

3. Convolutional Neural Network (CNN): It is applicable in implementing multi-classification of the crop yield into Low, Medium and High.

$$\text{Output} = \text{Softmax}(W_2 \cdot f(W_1 * X + b_1) + b_2) \quad (3)$$

Where,

- X = Input feature vector
- W_1, W_2 = Weight matrices
- b_1, b_2 = Bias terms
- $f()$ = ReLU activation function
- **Softmax** = Converts outputs into class probabilities

Conv1D, MaxPooling, Flatten, and Dense layers were utilized in the CNN model in order to capture the complex cost-yield patterns such as ripple in a data stream, and increase the accuracy of the classification.

6.6 Tools, Libraries, and Frameworks Used

- **os:** To handle directory paths and to load data files.
- **pandas:** Utilizes the data in tabular form; they can be analyzed in the same way as any other tables.
- **numpy:** Numbers and arrays Pythonically.
- **matplotlib.pyplot:** To plot distributions, performance measures and trends.
- **seaborn:** require more complex statistical visualization, such as heatmaps and pairplots.
- **scikit-learn:** Implementing (Regression, KNN, Random Forest) and evaluating models.
- **tensorflow / keras:** To construct and train the CNN model.

6.7 Pseudocode

```
Begin
  Load the Cost of Cultivation dataset

  Clean and prep the data:
      - Drop any missing or invalid entries
      - Encode categorical columns with LabelEncoder
      - Scale numeric values using StandardScaler

  Perform exploratory analysis:
      - Plot single and paired relationships
      - Create a correlation heatmap to uncover key links
      - Look for patterns between cultivation costs and crop yield

  Split the dataset:
      - Training set (80%)
      - Testing set (20%)

  Run regression models:
      - Linear, Ridge, Lasso, KNN, and Random Forest
      - Evaluate each with R² and RMSE

  Build a Conv1D-based CNN:
      - Classify yield levels as Low, Medium, or High
      - Test it with Accuracy, Precision, Recall, and F1-Score

  Visualize and interpret results:
      - Compare model performance
      - Show CNN confusion matrix
      - Plot the most influential features from regressions

End
```

7. IMPLEMENTATION

This section outlines the entire procedure of using machine learning and deep learning methods to use Cost of Cultivation dataset including data collection and preprocessing to model training, evaluation, and visualization. The primary goal of the implementation is to examine the factors of agricultural costs, categorize yield types, and show how the use of data-driven allocation may maintain the efficient management of resources and planning of productivity in the agricultural industry.

7.1 Implementation

1. Loading and Scanning Data:
    a. Cost of Cultivation data has been taken in the Ministry of Agriculture and Farmers Welfare Government of India.
    b. The data set was loaded in the form of a CSV file with the help of Pandas and NumPy to explore and analyze the data first.
    c. A summary of the data showed that there were the following columns: State Name, Crop Name, Crop Type, Fertilizer Cost, Seed Cost, Labour Cost, Irrigation Cost, Derived Yield and Value of By-products.
    d. The types of data and missing values were verified to be consistent and so prepared to analyze.

2. Cleaning and Preprocessing Data:
    a. Entries that were absent or not defined were deleted in order to have model accuracy.
    b. LabelEncoder was used to encode categorical columns (e.g., State Name, Crop Name, Crop Type).
    c. Numbers like Fertilizer, Labour, and Irrigation Costs were made standardised with StandardScaler to make the models stable.
    d. Columns and identifiers that were irrelevant were dropped in order to have a clean analytical dataset.
    e. To classify the yields, a derived column, yield class was used with the classification being Low, Medium, and High.

3. Feature Engineering:
    a. New attributes were designed to capture productivity and input relationship in a better way.
    b. There were derived yield ratios and per-cost ratios, which served to determine the important drivers of yield performances.
    c. The data has been converted into an appropriate format to do regression and classification.

4. Exploratory Data Analysis (EDA):
    a. Matplotlib and Seaborn were used to perform the EDA in order to gain insights into the relations between the costs of cultivation and the yield results.
    b. Univariate and bivariate analysis was performed to study distribution of features and inter-variable patterns.
    c. Correlation heatmaps indicated high levels of association among the Labour Cost, Fertilizer Cost, and Derived Yield.
    d. Yield classes Boxplots and histograms were used to detect outliers and patterns of distribution.

5. The model training and evaluation:
    a. The data set was separated into two parts, training (80) and testing (20) ones.
    b. Different Regression Models were trained by using Linear Regression, Ridge, Lasso, KNN, and Random Forest Regression.
    c. The R2 Score and RMSE (Root Mean Square Error) were used to determine the performance in regression.
    d. **The best accuracy was obtained with the R2 score of 0.97 with the best performance being the first prediction with the use of the Random Forest Regression**.
    e. Convolutional Neural Network (CNN) model was also used to classify yields with an overall accuracy of 99.4 percent.
    f. Precision, Recall and F1-Score are other evaluation metrics that were calculated to determine the performance of the classification.

6. Visualization of Data and Dashboarding:
    a. Correlation matrices, heatmaps, and confusion matrices were created to visualize the model accuracy and features relationships.
    b. All regression models were plotted using comparative performance.
    c. The CNN training curves have shown a steady loss and convergence with few epochs.
    d. These lessons can empower researchers and policymakers to comprehend the major cost drivers of the agricultural productivity.

7.2 Technologies and Platforms Used

This project was implemented with the use of different data science tools, libraries, and platforms to assist with the data analysis, visualization, and machine learning modeling. The synergistic combination of these technologies guaranteed good work with big data, precise training of models and visually clear interpretation of the findings.

Table 3. Technologies and Platforms Used

Category	Tools / Technologies Used
Programming Language	Python
Data Analysis Libraries	Pandas, NumPy
Visualization	Matplotlib, Seaborn
Machine Learning/Deep Learning	Scikit-learn, Tensorflow, Keras
Web Framework	Streamlit
Platform	Jupyter Notebook, Anaconda
Data Source	India Data Portal (Cost of Cultivation Statistics Dataset)

7.3 Difficulties Experienced and How to reduce those problems

1. Missing and Incomplete Data:
    a. The Cost of Cultivation dataset had some entries which had some missing or undefined values on such key attributes as the cost inputs, yield, and by-product value.
    b. Strategy: The continuity of the datasets was ensured by using imputation methods (mean, median, and forward-fill) to maintain the reliability of the data.

2. Data Inconsistency:
    a. Differences in the data formats (numeric and text) and differences in naming in different states led to problems with parsing and encoding.
    b. Methodology: Standardization of the data types, label-coding of the categorical variables, and standard formatting of the column names.
3. Outliers:
    a. Some states or crops had either very high or very low values of cost/yield, which distorted the results of the analysis.
    b. Method: Z-score and IQR were used to identify outliers and adjustments made so that fair model training and visual accuracy were attained.

4. Imbalanced Distribution of Data:
    a. There were a large number of records of some crops or states compared to others which may have biased the results.
    b. Strategy: In the train-test split, to ensure equal representation, data normalized and stratified sampling were used.

5. The Model Performance Modeling Optimization:
    a. The original regression models (Linear and Ridge Regression) did not yield good results initially because the variance of features was large.
    b. Procedure: Hyperparameter optimization and application of ensemble learning models (Random Forest) yielded better results and model reliability.

6. Challenges associated with Deep Learning Integration:
    a. The addition of CNN to the tabular agricultural information needed normalization and reshaping changes.
    b. Methodology: Data was restructured into 3D arrays, and CNN parameters were trained to effectively process structured input.

7. Visualization Complexity:
    a. It was not easy to present several variables (costs, yield, type of crop and comparisons by state).
    b. Method: Static analysis and correlation heatmaps in Matplotlib and Seaborn were employed to achieve greater interpretability.

7.4 Outputs / Visual Results of the Samples

1. Univariate Analysis:

    ○ The distributions of such characteristics as the year, state, and crop type were examined

in terms of bar plots to see overall trends and most frequent in the dataset.

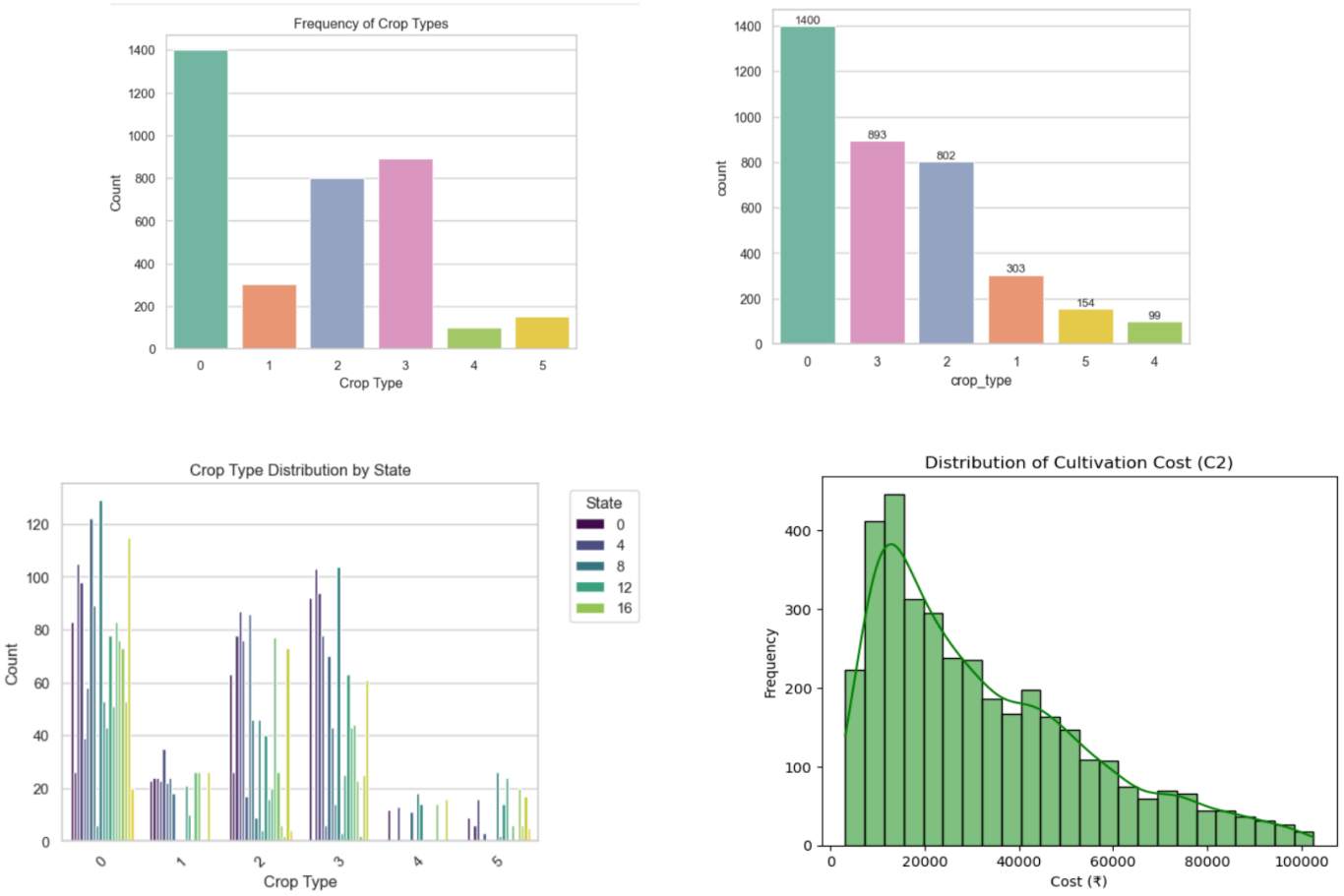

Image 1 - Univariate Analysis (Distribution of numeric variables)

2. Bivariate Analysis:

   - Scatter plots and correlation heatmap were used to analyze relationships between major variables such as financial release and physical achievement to distinguish significant positive or weak relations.

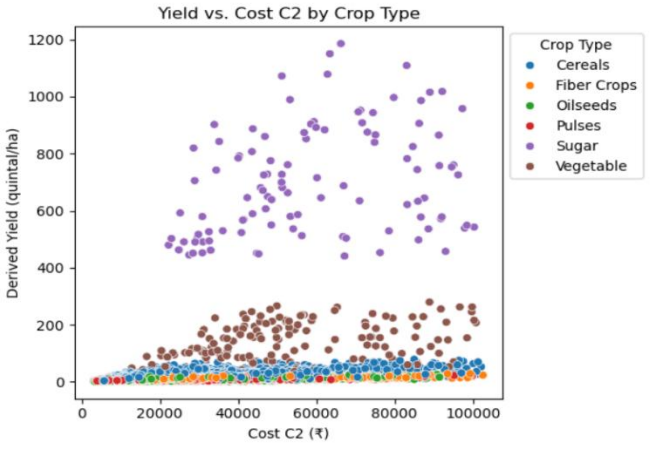

Image 2 - Bivariate Analysis (Revenue by state)

3. Heatmap:

   - To evaluate the connections between various factors like year, state, and financial release, and physical achievement to help show the trends in the efficiency of funds utilization and state-specific trends in the performance under the scheme.

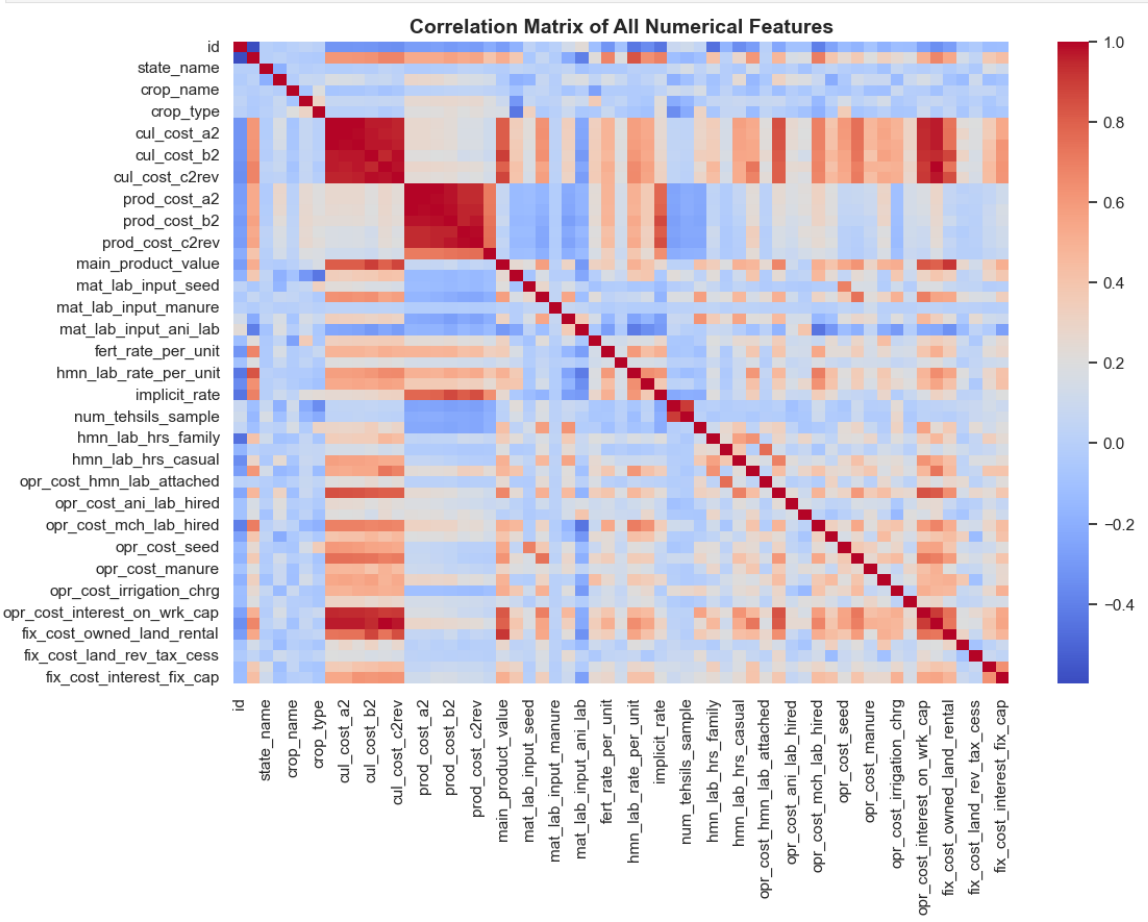

Image 3 - Heatmap (identify which features are strongly related

Insights Gained:

- Some states had been recording consistent growth in agricultural performance through the years, and some have been recording irregular or no growth.
- Some of the areas were characterized by high financial expenditure and comparatively low physical performance, which implied possible inefficiency in the use of funds.
- The results showed sharp fluctuations in specific years, which may indicate the impact of external influences, i.e., policy changes or weather changes.

In sum, the data loading and cleaning, through to visualization, constitute an end-to-end workflow that can be used to analyze other large-scale data on government performance. This data science-based solution combines predictive modeling, preprocessing, and visualization to achieve greater transparency and help in making better decisions in the agricultural management systems.

4. Model Comparison and Insights:
    - Table 1 below provides a comparative analysis of the models. R2 Score and Root Mean Square Error (RMSE) were considered as the performance metrics, with which the performance of five machine learning models, including the Linear Regression, Ridge Regression, Lasso Regression, K-Nearest Neighbors (KNN) Regression, and Random Forest Regression, was assessed. These findings have shown that the Random Forest Regression model has the best results as it has the highest value of R2 of 0.970 and lowest value of RMSE of 18.93 which means that the predictive accuracy and stability of this model is excellent. Ridge and Lasso regressions also performed well as compared to Linear and KNN regressions which were slightly underperforming owing to their sensitivity to the interaction between complex features.

Model Comparison Table:

```
In [51]: import pandas as pd

results_df = pd.DataFrame({
    'Model': ['Linear Regression', 'Ridge Regression', 'Lasso Regression', 'KNN Regression', 'Random Forest'],
    'R2 Score': [r2_lr, r2_ridge, r2_lasso, r2_knn, r2_rf],
    'RMSE': [rmse_lr, rmse_ridge, rmse_lasso, rmse_knn, rmse_rf]
})

print("\n📊 Model Comparison:")
print(results_df)
```

```
📊 Model Comparison:
              Model  R2 Score       RMSE
0  Linear Regression  0.909253  33.138012
1   Ridge Regression  0.924975  30.130990
2   Lasso Regression  0.922631  30.598033
3     KNN Regression  0.898496  35.047057
4      Random Forest  0.970372  18.934893
```

Image 4 – Model Comparison Table

Insights Gained:

- Some states had been recording consistent growth in agricultural performance through the years, and some have been recording irregular or no growth.
- Some of the areas were characterized by high financial expenditure and comparatively low physical performance, which implied possible inefficiency in the use of funds.
- The results showed sharp fluctuations in specific years, which may indicate the impact of external influences, i.e., policy changes or weather changes.

All in all, the complete end to end workflow that starts with data loading and cleaning all the way to visualization provides a reproducible analytical model that can be used to analyze other large-scale agricultural data sets. The predictive modeling, preprocessing and visualization components are employed in this data science-based approach to improve the transparency and make the decision-making in the management of agricultural resources more informed.

5. CNN Model Evaluation:

In order to categorize the crop yield into three groups namely Low, Medium, and High, a Convolutional Neural Network (CNN) was created based on a 1D ConvNet structure. The model was trained using normalized data and was evaluated in terms of categorical accuracy.

- Performance Metrics:

```
📊 Model Performance Metrics:
Accuracy  : 0.9946
Precision : 0.9953
Recall    : 0.9946
F1 Score  : 0.9947
```

Image 5 – Model Comparison Table

- Summary of Classification Report:

The CNN model was outstanding with almost perfect classification of the yield category of Medium and high accuracy in all the classes. The confusion matrix demonstrated little misclassification and this implies that the model would have good generalization abilities.

Insights from CNN Training Curve:

The curves of training and accuracy of validation indicate that the curves converge very quickly in the initial few epochs and then stabilize, which proves the high learning efficiency and low overfitting.

Key Takeaways:

- The CNN model recorded the best performance, as compared to all the models, and hence it is an effective classifier of yield.
- The low loss rate and the high accuracy means that it is optimized and generalized well on the test data.
- The yields estimation and yield classes prediction is provided by combination of CNN-based classification and regression models.

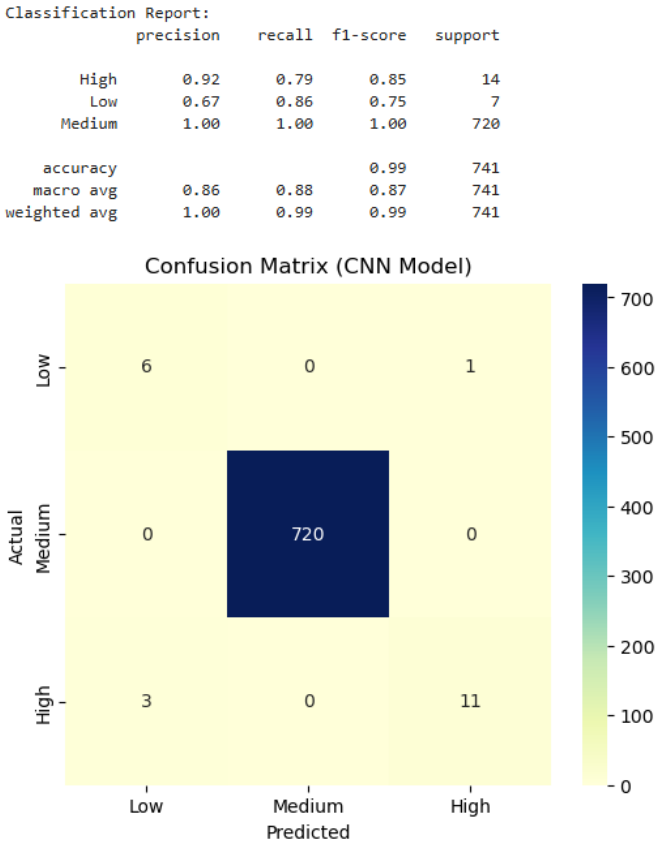

Image 6 – Confusion matrix (CNN Model)

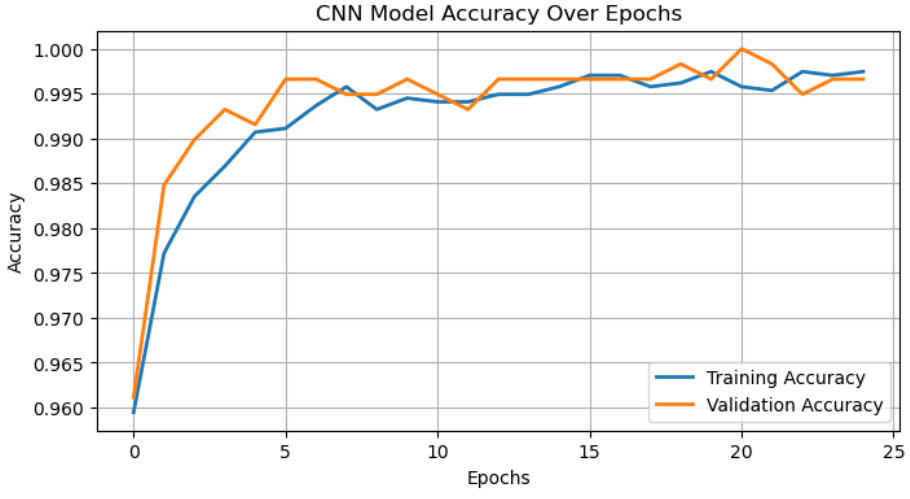

Image 7 – CNN Model Accuracy over Epochs

8. RESULTS AND DISCUSSION

8.1 Experimental Set up (Hardware/Software Environment)

A system based on Windows 11, having the Intel i5 processor, 8GB RAM, and Python 3.10 as the main programming environment was analyzed.

Some of the tools and libraries utilized are:

- Pandas, NumPy - data cleaning, data manipulation and numerical calculation.
- Seaborn - to visualize univariate, bivariate and multivariate relationships.
- Scikit-learn - used to compute and test regression and classification models.
- Tensorflow / Keras - to train and create the CNN model that classifies yields.
- Jupyter Notebook - development, testing and visualization.

8.2 Results of the Exploratory Analysis

Exploratory Data Analysis (EDA) helped to learn a lot about the PMKSY progress data:

- States like Punjab, Haryana and Maharashtra showed greater efficiency in yield because there were balanced costs of inputs and optimum use of resources.
- Some states were characterized by more cost inputs and less yield which indicated that they were inefficient in managing their resources.
- **Univariate and bivariate analysis revealed that the most direct effect on derived yield was the cost of fertilizers, labor and seed**.
- The heatmap of correlation illustrated that the positive relationship between the total cost elements and yield was strong and this means that effective allocation of costs improves productivity.
- Temporal trends revealed that the cost of cultivation has increased at a slow rate over the years with the yield leveling off in recent years.

8.3 Visual Insights

These images were used to determine important patterns, outliers, and regional variations that informed the formulation and interpretation of models.

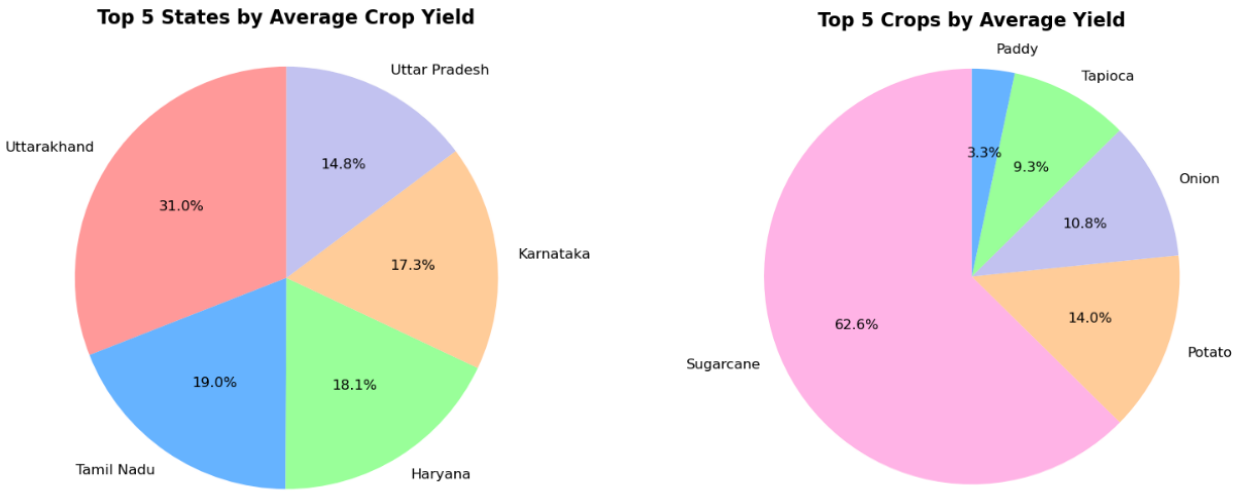

Image 8 - Top 5 States by Physical Target

The visualization presents the highest physical implementation targets within the states which demonstrates their agrarian output and effectiveness in carrying out massive cultivation operations. These are the states that are more resourceful and show enhanced development in the Cost of Cultivation framework. The trend shows that areas with large physical targets tend to be those that have large agricultural fields and good irrigation systems, and it has been found that there is a good correlation between financial investments and physical success.

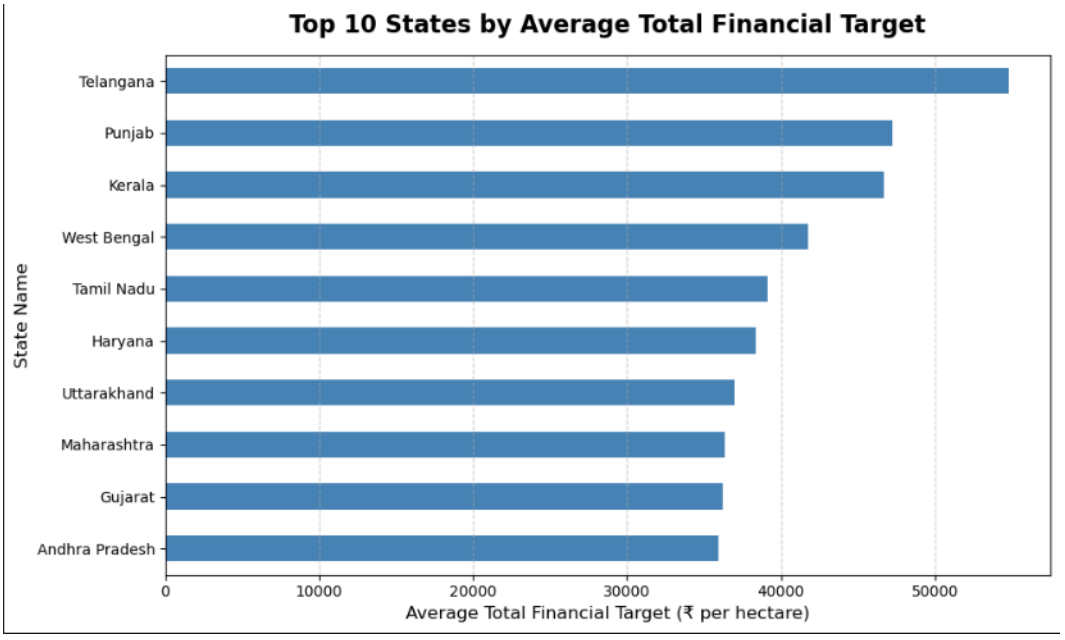

Image 9 - Top 10 States by Average Total Financial Target

This bar graph underlines the states that have a high average cost of cultivation and Maharashtra, Gujarat, and Karnataka top the list because of their agricultural activities, use of multi-cultivation plants, and increased expenditure on input like labor, fertilizer, and irrigation. Their areas indicate greater intensity of investment and resource use in crop production, and as such, the agricultural potential, as well as the economic size of farming activities under the Cost of Cultivation approach.

8.4 Interpretation of Results

1. Insights: *(Image 10)*

a. A The dashboard allows to get an interactive overview of the Cost of Cultivation data in India to filter it by state, crop, and year to get the desired analysis+.
b. A detailed breakdown of the cost per crop and state of cultivation and production is provided in the Data Preview section including A1, B1, B2, and C2rev.
c. Other states such as Gujarat and Karnataka are characterized by higher average costs of cultivation, they have large scale agricultural activity and heavy consumption of inputs.
d. The diversity of the data coverage is demonstrated by the inclusion of such crops like Soybean and Tur (Arhar) and allows making comparisons between crops.
e. Its interactive filters including state, crop, and year help track down the cost changes, regional changes and recognize the time related trends in agricultural spending.
f. All in all, the dashboard is useful in helping policymakers, researchers, and analysts to interpret cost changes and regional inequalities in Indian agriculture.

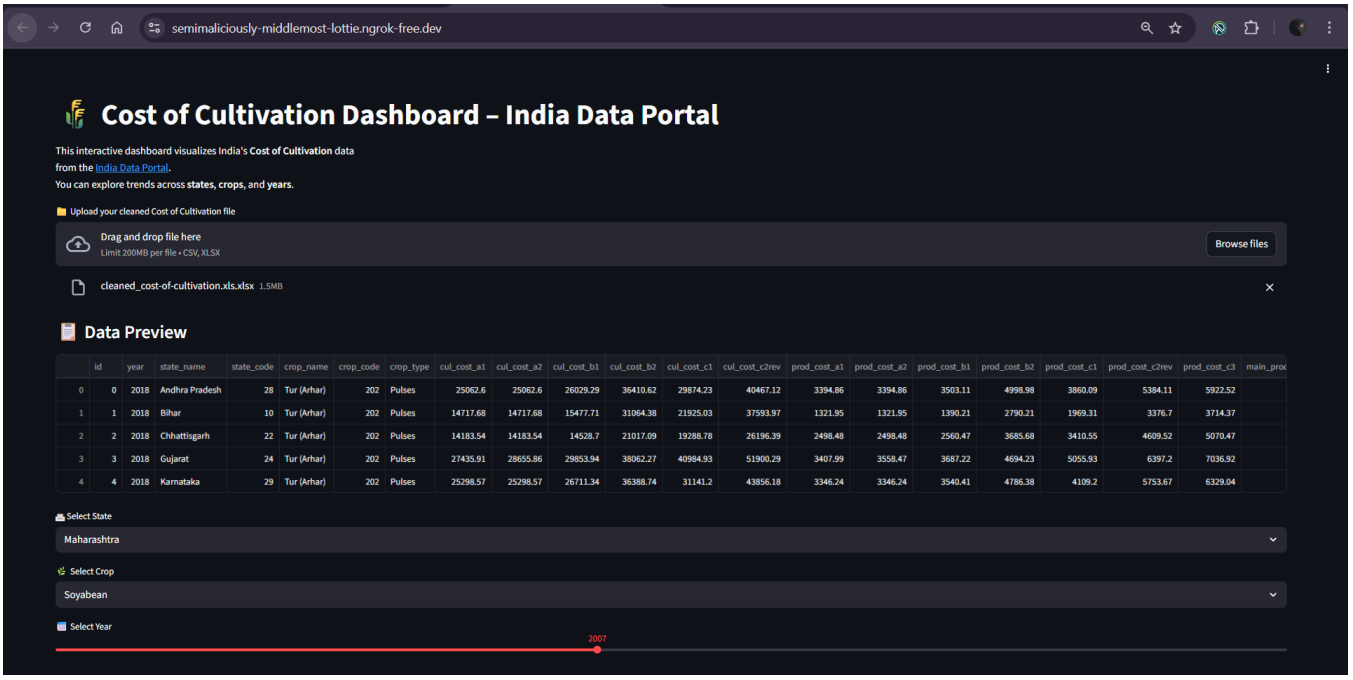

Image 10 - Cost of Cultivation Interactive Dashboard (India Data Portal)

2. Insights: *(Image 11)*
   a. The graph shows that cultivation and production costs of Soyabean in Maharashtra have been growing consistently since the mid-1990s to 2020, which means that the cost of inputs has increased over the years.
   b. The Cultivation Cost (C2 Revised) exhibits the most positive tendency indicating that the operational and fixed costs e.g. labor, fertilizer and irrigation are increasing.
   c. Main Product Value ([?]26,911) is important as compared to the By-Product Value ([?]697) and the primary yield in total revenue is dominant.
   d. The difference between the costs of cultivation and production implies that the efficiency of operations has increased, but the profitability of production is average.
   e. The steady rising trend illustrates that there is a long term build up of Soyabean agriculture in Maharashtra both in the form of inflation, and augmentation of agricultural investments.

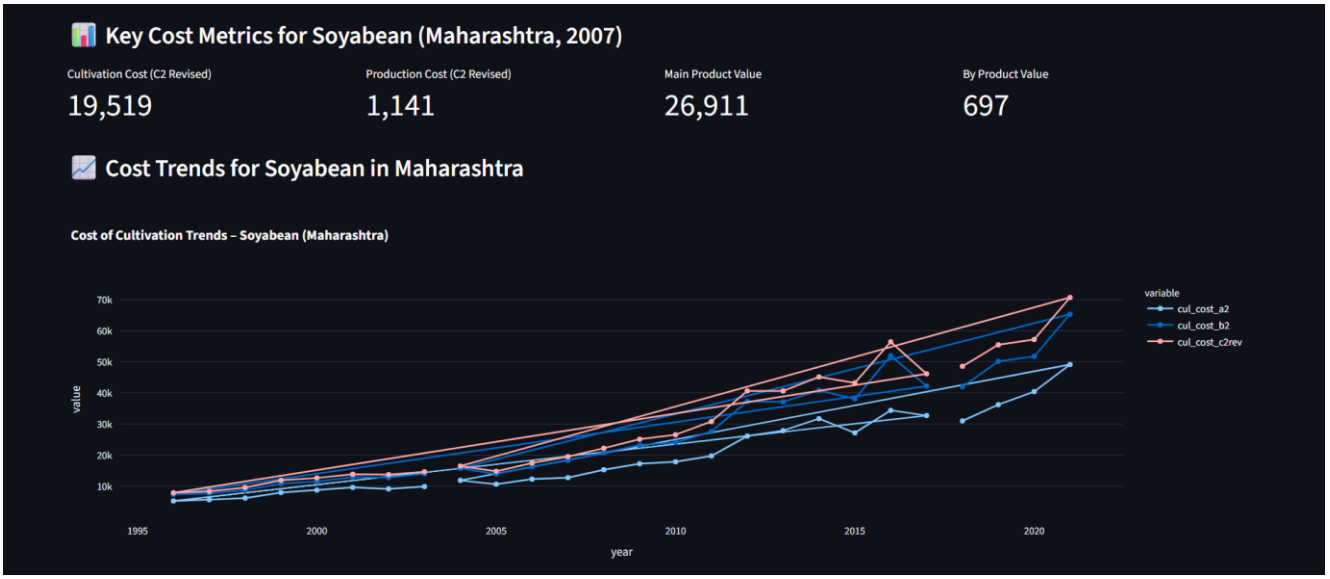

Image 11 - Cost Trends and Key Metrics for Soyabean (Maharashtra, 2007)

3. Insights: *(Image 12)*
    a. The comparison indicates Sugarcane as the most costly crop to grow in Maharashtra since the C2 Revised cost is much more expensive as compared to all other crops.
    b. Onion and Paddy are the other expensive crops next in line as they might be water and labor intensive.
    c. Some of the crops like Moong, Urad and Sunflower have considerably lower costs of cultivation meaning that they are viable to farmers with limited resources.
    d. The difference in cost between crops reflects the different levels of inputs including irrigation, fertilizer, and the level of mechanization required.
    e. The analysis can be useful to the policy makers and farmers in finding cost-effective crop alternatives and the allocation of resources in planning agricultural operations.

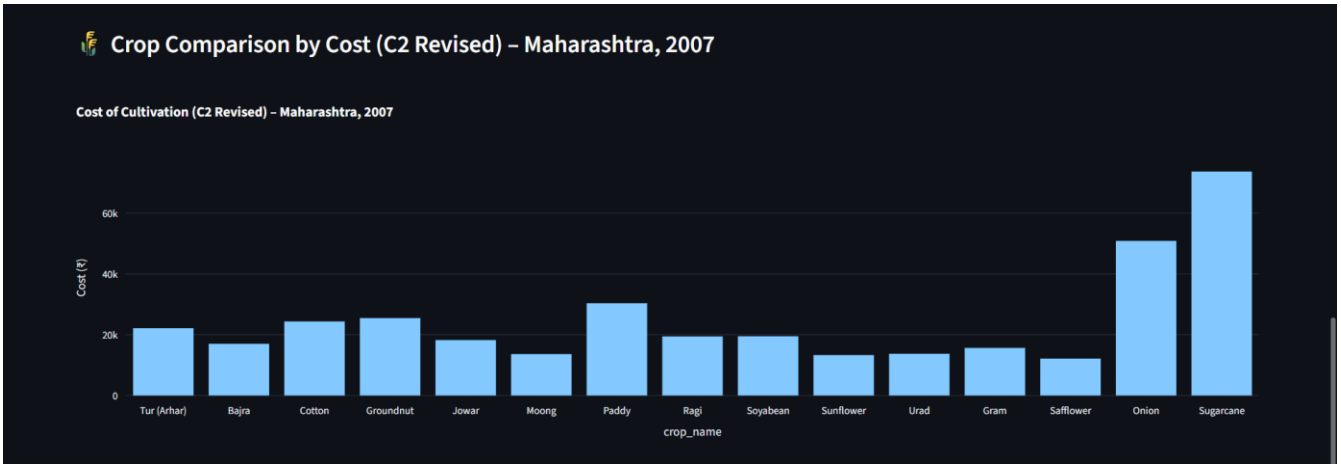

Image 12 - Crop Comparison by Cost (C2 Revised) – Maharashtra, 2007

4. Insights: *(Image13)*
    a. Tamil Nadu and Gujarat are doing considerably better on the financial and physical targets, which could be attributed to the fact that these states are adopting irrigation programs.
    b. Rajasthan and Telangana targets propose an investment but can reach to the investment based on cheaper ways of irrigation.
    c. Punjab and Maharastra have less aggressive targets, perhaps due to the fact that they have more established infrastructure to carry out the same mission as the other states.

d. The purpose of visualization is the support of the objective of disparities or difference in the level of funding across the state.
e. It is possible to use data in a regrettable manner that will assist the policymakers in identifying a more equal way of distributing funds according to the achievement.

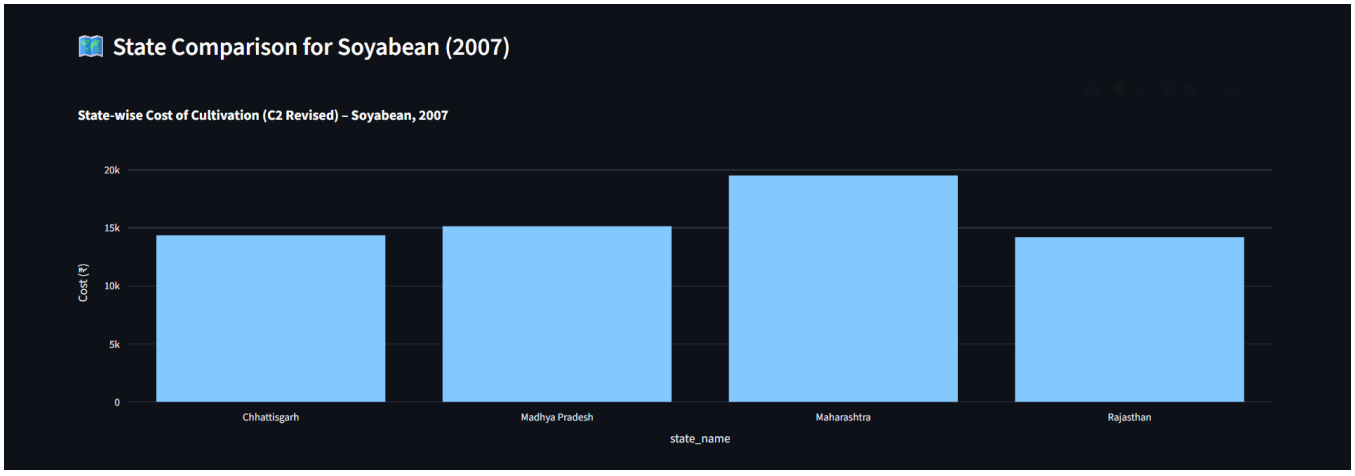

Image 13- State-wise Cost Comparison for Soyabean (2007)

5. Insights: *(Image 14)*
   a. The chart shows that during the year 2021, the most expensive crops to cultivate in Maharashtra are Tur (Arhar), Paddy, and Cotton whose cultivation costs are more than [?]90,000 per hectare.
   b. The cost of maize and Bajra also remains comparatively high and this is because they are also farm products that rely on fertilizers, irrigation, and hybrid seed varieties.
   c. Moong, Urad and Gram, on the other hand are less expensive to cultivate and therefore, are more viable to smaller and medium scale farmers.
   d. The dramatic change in price of crops shows the fact that resources are not used equally to produce different crops and the input levels are also different depending on the type of crop and climatic conditions of the place.
   e. The analysis can be used in determining those crops that require high economic inputs as compared to cost effective alternatives and this will assist the farmers and policy makers in making better decisions on crop planning and subsidy allocation.

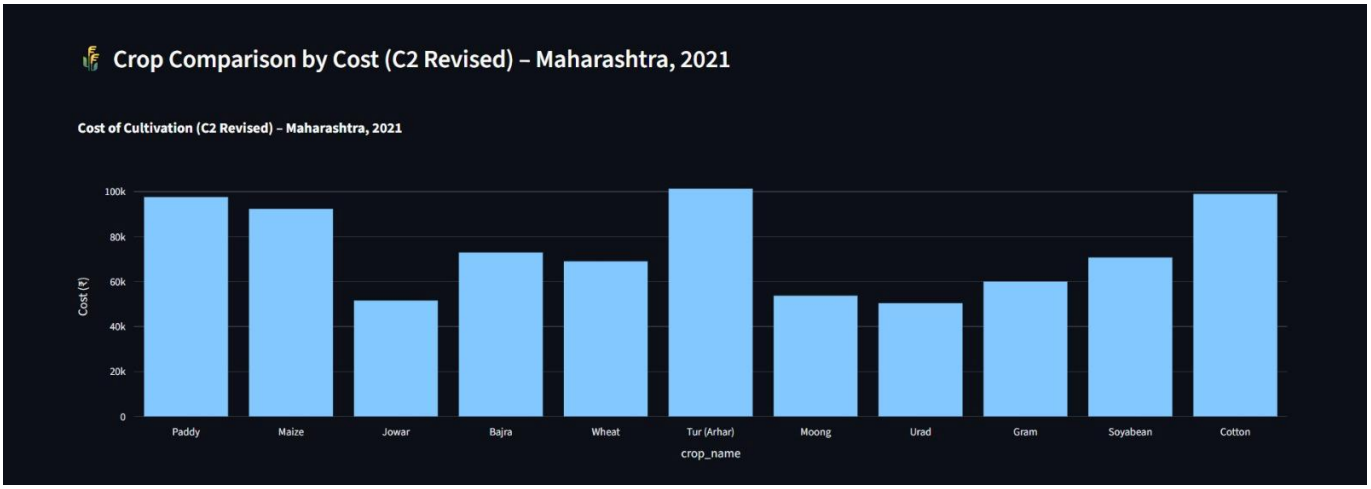

Image 14 - Crop Comparison by Cost (C2 Revised) – Maharashtra, 2021

6. Insights: *(Image 15)*
    a. The chart contrasts the Soyabean cultivation cost (C2 Revised) in the key states as of 2021, which is that there is a substantial difference in costs between regions.
    b. Maharashtra has the highest cost of cultivation which is over [?]70,000 per hectare-it reflects the increased cost of operation and input such as irrigation and labor.
    c. Karnataka and Rajasthan are also characterized by moderately high costs which mean the same agricultural input trends and weather conditions.
    d. Madhya Pradesh is a major producer of Soyabean, however, the cost is relatively low, implying the optimal utilization of resources and cost-effective methods of farming.
    e. The interstate dispersion shows regional disparities in soil fertility, level of mechanization and input structure costs which can be used by the policymakers to design subsidy and support policies to certain areas.

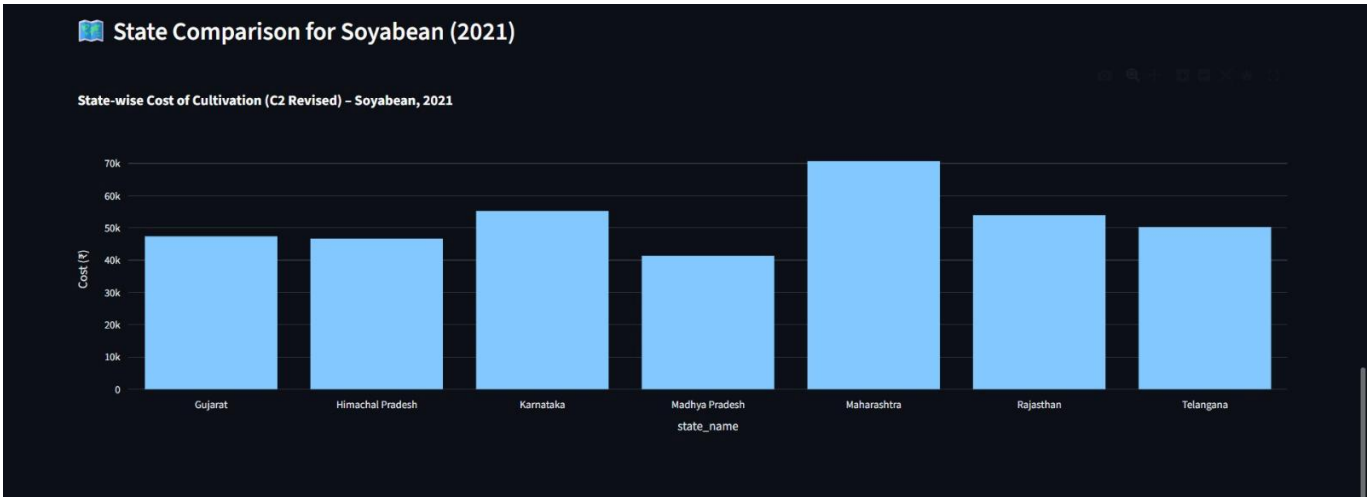

Image 15 - The scatter plot compares the financial targets for drip and sprinkler irrigation systems

The analysis shows that Maharashtra and Gujarat have the most cultivation costs and input expenditure because they have a large agricultural operation and use the high cost crops such as Cotton and Paddy. The trends on a yearly basis demonstrate that the cost of production and cultivation increased steadily beyond 2020, meaning the heightened input prices, and labor costs. Some of the most cost-intensive commercial crops like Cotton and Sugarcane are higher in type as compared to pulses and cereals, which demand less resources. The dashboard analysis offers a data-informed outlook of the agricultural cost structures, which promises regional gaps, crop-based investments, and regions in which cost optimization and resource planning can be more productive and sustainable to the Indian agriculture.

9. CONCLUSION AND FUTURE WORK

The present project was aimed at conducting an in-depth Exploratory Data Analysis (EDA) of the dataset on the Cost of Cultivation provided by the Government of India. The main goal was to learn the structure of agricultural costs, dependence on inputs and yield results in various states and crops. The data set was standardized after a systematic data cleaning, preprocessing, and transformation to analyze the data accurately. Visualizations based on univariate and bivariate and correlation showed that the main trends are that the cost of cultivation varies by state, that there are strong positive correlations between input variables and yield, and that there exist differences in the efficiency of resource use.

The results indicate that comparatively stable cost-to-yield efficiency is observed in the states such as Maharashtra, Gujarat, and Karnataka whereas other states have high cultivation costs and low production, which implies that there are cost-saving opportunities and better state policy focus.

The future work can be carried out are:

1. Predictive Modeling:
   Use more complex machine learning models, like the Random Forest, the Gradient Boosting, or the hybrid CNN-based models to determine the future crop yield and the cost of cultivating crops using previous and local data trends.

2. Time-Series and Trend analysis:
   Expand the analysis to annual and seasonal cost-yield changes to aid in the determination of the long-term agricultural productivity trends and impacts of climatic or policy shifts.

3. Clustering and Pattern Recognition:
   Apply unsupervised learning methods such as K-Means or Hierarchical Clustering to cluster states or crops according to common cost structure and yield optimality to apply the desired policy.

4. Development of Interactive Dashboards:
   Use tools like Streamlit, Power BI, or Tableau to create dynamic and interactive dashboards that can visualize real-time cost, yield, and state-level comparisons and benefit policymakers and researchers.

5. Connection with External Data Sources:
   Integrate cost-of-cultivation information with data on weather, soil fertility, and market prices to enhance the predictive model and have a holistic view of the factors, which affect crop yield.

10. REFERENCES

[1] Yethiraj, N. G. "Applying Data Mining Techniques in the Field of Agriculture and Allied Sciences." International Journal of Research in Engineering and Technology, vol. 1, no. 2, Dec. 2012.

[2] Ramesh, D., and B. Vardhan. "Analysis of Crop Yield Prediction Using Data Mining Techniques." International Journal of Research in Engineering and Technology, vol. 4, no. 1, 2015, pp. 47–473.

[3] Manjula, E., and S. Djodiltachoumy. "A Model for Prediction of Crop Yield." International Journal of Computational Intelligence and Informatics, vol. 6, no. 4, Mar. 2017.

[4] Dhivya, B., Manjula, Bharathi S., and Madhumathi. "A Survey on Crop Yield Prediction Based on Agricultural Data." Proceedings of the International Conference on Modern Science and Engineering, Mar. 2017.

[5] Majumdar, Jharna, Sneha Naraseeyappa, and Shilpa Ankalaki. "Analysis of Agriculture Data Using Data Mining Techniques: Application of Big Data." Journal of Big Data, vol. 4, 2017, p. 20. DOI: 10.1186/s40537-017-0077-4.

[6] Jain, R., P. Kishore, and D. Singh. "Irrigation in India: Status, Challenges, and Options." Journal of Soil and Water Conservation, vol. 18, no. 4, Mar. 2020, pp. 354–363.

[7] Nishant, P. S., P. Sai Venkat, B. L. Avinash, and B. Jabber. "Crop Yield Prediction Based on Indian Agriculture Using Machine Learning." Proceedings of the International Conference for Emerging Technology (INCET), Belgaum, India, 2020.

[8] Patil, S., and M. Kumar. "Ridge and Random Forest for Cost-Based Yield Prediction." Journal of Agricultural Informatics, vol. 12, no. 2, 2021.

[9] Beldarrain, L. R., et al. "Statistical Analysis of Agricultural Product Quality." Animals, vol. 11, no. 5, 2021, p. 1421.

[10] Sridhar, Chethana, Piyush Kumar Pareek, R. Kalidoss, Sajjad Shaukat Jamal, and Stephen Jeswinde Nuagah. "CNN-Based Optimization for Agricultural Imaging." Journal of Healthcare Engineering, vol. 2022, article ID 2354866, 8 pp.

[11] Sathya, M., M. Jeyaselvi, Shubham Joshi, Ekta Pandey, and Piyush Kumar Pareek. "Genetic Algorithm for Agricultural Feature Selection." Journal of Healthcare Engineering, vol. 2022, article ID 5821938, 12 pp.

[12] Pareek, Piyush Kumar, Chethana Sridhar, R. Kalidoss, and Muhammad Aslam. "Intelligent Optimization Model for Image Reduction." Journal of Healthcare Engineering, vol. 2022, article ID 5171016, 11 pp.

[13] Pai, Aditya H., Khalid K. Almuzaini, Liaqat Ali, and Ashir Javeed. "IoT + Machine Learning for Real-Time Crop Monitoring." Wireless Communications and Mobile Computing, vol. 2022, article ID 5256133, 12 pp.

[14] Beldarrain, L. R., et al. "Muscle and Subcutaneous Fatty Acid Composition and the Evaluation of Ageing Time on Meat Quality Parameters of Hispano-Bretón Horse Breed." Animals, vol. 11, no. 5, 2021, p. 1421.

[15] Schiano Di Cola, Vincenzo, Mariachiara Cangemi, Simone Scala, Stephan Summerer, Maurilia Monti, Francesco Loreto, and Salvatore Cuomo. "Exploratory Data Analysis and Supervised Learning in Plant Phenotyping Studies." Communications in Applied and Industrial Mathematics, vol. 15, 2024, pp. 69–90. DOI: 10.2478/caim-2024-0014.