SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR
Symbiosis International (Deemed University)
(Established under section 3 of the UGC Act, 1956)
Re-accredited by NAAC with 'A++' grade | Awarded Category – I by UGC
Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

A PROJECT REPORT

ON

"Indian Crop Cost & Profitability Analysis Dashboard"

Submitted By
Ruthvek Kannan (22070521031)

UNDER THE GUIDANCE OF

Dr. Piyush Chauhan
Associate Professor (Examination Incharge SIT Nagpur)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

AY 2025-26

Abstract

The Indian agricultural sector, a cornerstone of the nation's economy, faces persistent challenges in maintaining profitability against a backdrop of rising and volatile cultivation costs. Effective decision-making for farmers, policymakers, and agricultural analysts is contingent upon a deep and accessible understanding of these cost structures. This project addresses this critical need by designing, developing, and evaluating a comprehensive, interactive data science dashboard built with Streamlit.

This report documents the complete lifecycle of the project, from data ingestion to the deployment of a multi-functional web application. We utilized the "Cost of Cultivation of Principal Crops in India" dataset to perform a rigorous Exploratory Data Analysis (EDA), revealing significant trends such as the "cost-price squeeze" affecting farmer profitability.

The core of the project involved a **three-pronged modeling strategy** to analyze the data from every angle:

1. **Regression:** To predict exact production costs per quintal.
2. **Classification:** To categorize farms into cost efficiency tiers.
3. **Clustering:** To discover unsupervised "farm archetypes" based on operational data.

Using a robust **5-fold cross-validation** strategy, we identified **XGBoost Regressor** as the optimal model for cost prediction (Mean CV R^2: **0.89**) and **Support Vector Classifier (SVC)** for efficiency classification (Mean CV Accuracy: **0.65**). These models were then integrated into an interactive **Streamlit** dashboard. The final application empowers users with advanced tools, including a **Future Cost Predictor**, a **"What-If" Scenario Analyzer**, and a **Profitability Calculator**, bridging the gap between complex data science and actionable agricultural insights.

Table of Contents

1. **Introduction**
    1.1. Project Background and Problem Statement
    1.2. Project Objectives (Aim)
    1.3. Scope and Limitations
2. **System Design and Methodology**
    2.1. Technology Stack
    2.2. System Architecture & Data Flow
    2.3. Data Ingestion and Preprocessing
    2.4. Feature Engineering
3. **Exploratory Data Analysis (EDA)**
    3.1. Key Findings from Visual Analysis
4. **Modeling Methodology and Evaluation**
    4.1. Task 1: Regression (Cost Prediction)
    4.2. Task 2: Classification (Efficiency Categorization)
    4.3. Task 3: Clustering (Farm Archetype Discovery)
5. **Interactive Dashboard Features (The Application)**
    5.1. Predict Future Crop Cultivation Cost
    5.2. "What-If" Scenario Analysis
    5.3. Profitability Calculator
6. **Conclusion**
7. **Future Work**

1. Introduction

1.1. Project Background and Problem Statement

Agriculture is the backbone of the Indian economy, supporting millions of livelihoods. Despite its significance, the sector is fraught with financial uncertainty. A primary driver of this uncertainty is the high and increasingly volatile cost of cultivation. Farmers face a "cost-price squeeze" where the prices of inputs like fertilizer, labor, and machinery rise steadily, while the market prices for their crops remain volatile or stagnant.

This creates a critical need for data-driven tools. Policymakers require accurate cost data to set effective Minimum Support Prices (MSPs). Farmers need to budget for future seasons and understand their financial risks. Agricultural analysts need to identify regional inefficiencies and opportunities for improvement.

While the government collects extensive data through the "Comprehensive Scheme for studying Cost of Cultivation," this data is often stored in raw, inaccessible formats. The problem this project solves is the gap between the existence of this valuable data and its practical, actionable use by end-users. Without accessible tools, this data remains underutilized, and stakeholders are forced to rely on anecdotal evidence rather than data-driven insights.

1.2. Project Objectives (Aim)

The primary aim of this project is to bridge this gap by designing, building, and deploying a comprehensive, interactive data science dashboard to analyze and predict Indian crop cultivation costs.

The specific objectives are:

1. **To Collect and Prepare** the "Cost of Cultivation of Principal Crops in India" dataset from data.gov.in.
2. **To Perform** a thorough Exploratory Data Analysis (EDA) to visualize key trends, distributions, and relationships in cost, yield, and profitability.
3. **To Implement and Evaluate** a suite of machine learning models for three distinct tasks:

- **Regression:** To predict the *exact* production cost (prod_cost_c2).
- **Classification:** To categorize farms by *cost efficiency* relative to their peers.
- **Clustering:** To discover natural *farm archetypes* based on operational profiles.

4. **To Deploy** these models into an interactive, user-friendly Streamlit application featuring advanced tools for prediction, scenario planning, and profitability calculation.

1.3. Scope and Limitations

This project covers the complete data science workflow, from data ingestion to a functional web application. The analysis and models are based solely on the features present in the provided dataset.

Limitations:

- The analysis is constrained by the available features. It does not include external data like micro-climate weather data, soil quality information, or real-time market prices, which could further enhance prediction accuracy.
- The prediction models, while robust, are based on historical trends. The trend extrapolation for future cost prediction is a linear approximation and cannot account for sudden, unforeseen events (e.g., droughts, policy changes, global supply chain disruptions).
- The SVR and SVC models, while functional, have not undergone exhaustive hyperparameter tuning (like GridSearchCV) to prioritize application speed.

2. System Design and Methodology

2.1. Technology Stack

The project was developed using the Python ecosystem, leveraging a modern stack for data science and web deployment.

- **Web Framework:** Streamlit was chosen for its ability to rapidly build and deploy beautiful, interactive data applications purely in Python.
- **Data Manipulation:** pandas and numpy were used for all data loading, cleaning, transformation, and feature engineering tasks.
- **Machine Learning:** scikit-learn was the primary library for building preprocessing pipelines (Pipeline, ColumnTransformer), feature scaling (StandardScaler), encoding (OneHotEncoder), and implementing a wide range of models (Linear Regression, SVR, SVC, K-Means, etc.). xgboost was used for its high-performance gradient boosting models.
- **Data Visualization:** matplotlib and seaborn were used to generate the static, publication-quality plots for EDA and model evaluation. plotly was used for the interactive charts (bar chart, waterfall chart) in the advanced user tools.

2.2. System Architecture & Data Flow

The application is a multi-page dashboard where navigation is controlled by a main sidebar. The logic is modularized into separate Python files to adhere to the Single Responsibility Principle and improve maintainability.

1. **app.py:** The main entry point. It handles page routing, loads the data, and calls the appropriate functions from other modules based on user navigation.
2. **eda.py:** A dedicated module containing the perform_eda function. It receives the loaded DataFrame and generates all EDA visualizations.
3. **models.py:** The core analytical module. It contains all functions for data preprocessing (preprocess_data), model training and evaluation (e.g., run_regression_models), and the logic for all interactive tools (e.g., predict_cost, run_what_if_analysis).

Application Architecture Flowchart

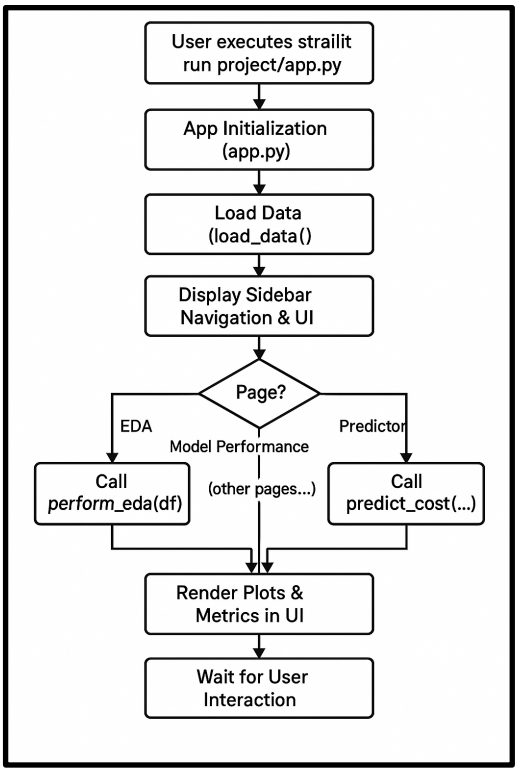


Figure 1. High-Level System Architecture Flowchart

2.3. Data Ingestion and Preprocessing

Data is loaded once at the start of the app using the load_data function in app.py. This function is wrapped in an in-memory caching decorator to ensure the CSV is only read from disk once per session.

This initial load is followed by preprocessing steps defined in models.py's preprocess_data function. This function is called by each modeling and analysis function to ensure data is correctly prepared.

1. **Missing Value Imputation:** fillna(df.median()) is used to fill missing numerical values with the median of their respective columns.
2. **Feature Scaling:** StandardScaler is used within a scikit-learn pipeline. This scales all numeric features to have a mean of 0 and a standard deviation of 1.
3. **Categorical Encoding:** OneHotEncoder (with handle_unknown='ignore') is used within the pipeline to convert categorical features like state_name and crop_name into a binary vector format.

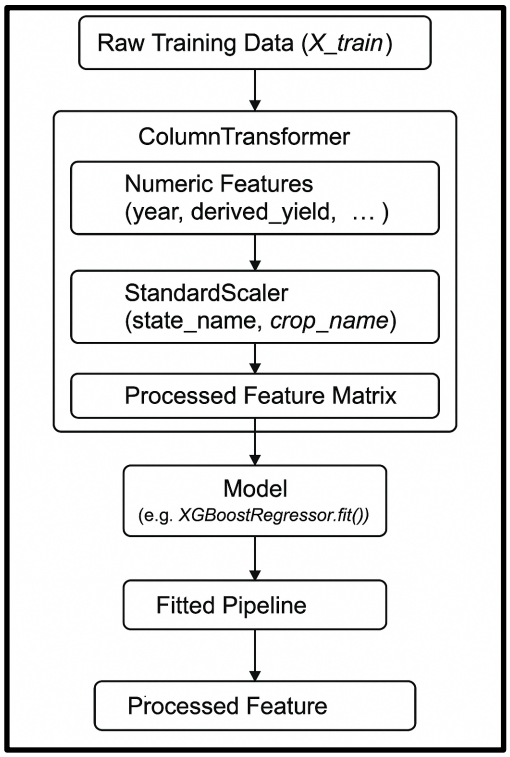


Figure 2.Data Preprocessing and Modeling Pipeline

2.4. Feature Engineering

Feature engineering was critical for this project. New features were created to provide more predictive power and to enable meaningful clustering.

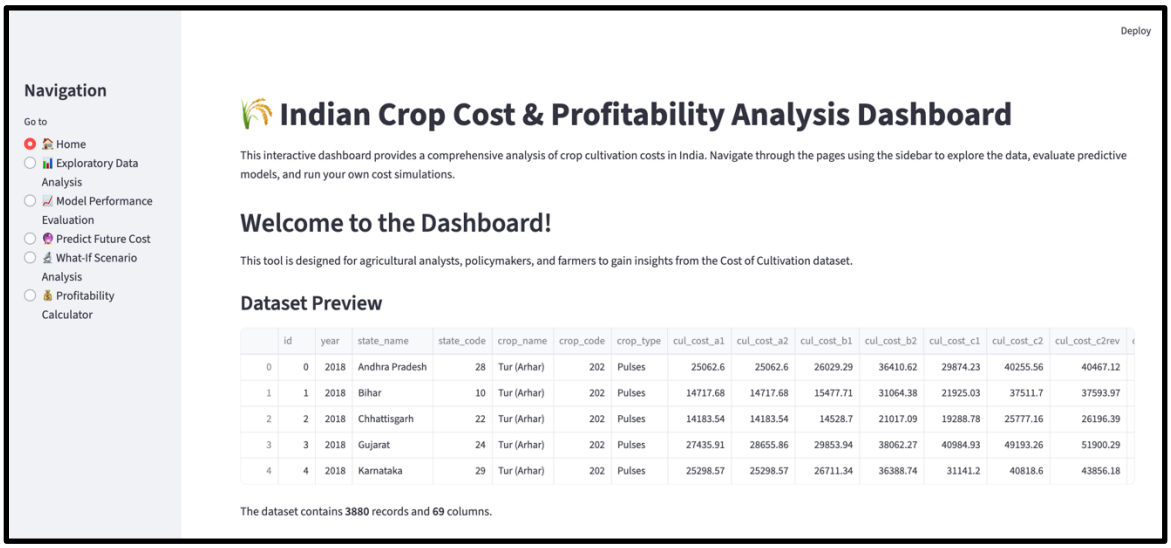
Table of Engineered Features

Feature Name	Formula	Purpose
General Features		
Total Human Labor Cost	Sum of all human labor cost columns	Aggregates human labor costs into one predictor.
Total Machine Labor Cost	Sum of all machine-related cost columns	Aggregates machine operation costs into one predictor.
Net Return	(Main Product Value + By-Product Value) − Total Cultivation Cost	Measures profitability; used in EDA and clustering.
Clustering Features		
Labor Cost Ratio	Total Human Labor Cost ÷ Total Cultivation Cost	Measures reliance on manual labor (labor intensity).
Machine Cost Ratio	Total Machine Labor Cost ÷ Total Cultivation Cost	Measures reliance on machinery (capital intensity).

Fertilizer Cost Ratio	Fertilizer Cost ÷ Total Cultivation Cost	Measures dependence on chemical inputs.
Mechanization Index	(Machine Labor Cost + 1) ÷ (Human Labor Cost + 1)	Compares machine vs. labor usage; avoids division-by-zero.

3. Exploratory Data Analysis (EDA)

The "📊 Exploratory Data Analysis" page of the dashboard presents a detailed visual analysis of the dataset, generated by eda.py. The "🏠 Home" page provides a quick preview of the loaded data.



	id	year	state_name	state_code	crop_name	crop_code	crop_type	cul_cost_a1	cul_cost_a2	cul_cost_b1	cul_cost_b2	cul_cost_c1	cul_cost_c2	cul_cost_c2rev
0	0	2018	Andhra Pradesh	28	Tur (Arhar)	202	Pulses	25062.6	25062.6	26029.29	36410.62	29874.23	40255.56	40467.12
1	1	2018	Bihar	10	Tur (Arhar)	202	Pulses	14717.68	14717.68	15477.71	31064.38	21925.03	37511.7	37593.97
2	2	2018	Chhattisgarh	22	Tur (Arhar)	202	Pulses	14183.54	14183.54	14528.7	21017.09	19288.78	25777.16	26196.39
3	3	2018	Gujarat	24	Tur (Arhar)	202	Pulses	27435.91	28655.86	29853.94	38062.27	40984.93	49193.26	51900.29
4	4	2018	Karnataka	29	Tur (Arhar)	202	Pulses	25298.57	25298.57	26711.34	36388.74	31141.2	40818.6	43856.18

Figure 3.Project Dashboard Home Page

3.1. Key Findings from Visual Analysis

- **Cost Drivers:** The "Top 10 Costliest Crops" chart shows that Sugarcane and Onion are the most capital-intensive crops, often costing over **95,000** per hectare.
- **Profitability:** The "Distribution of Net Return" histogram reveals that a significant number of farming operations run at a financial loss (left of the break-even line).
- **Temporal Trends:** The most significant finding is the **"cost-price squeeze"** visualized in Fig 4. While cultivation costs (cul_cost_c2) have risen steadily, net returns (net_return) have remained volatile and stagnant, highlighting the immense financial pressure on farmers.
- **Correlations:** The "Correlation Matrix of Key Variables" (Fig 5) confirms that total_human_labor_cost and opr_cost_fertilizer are two of the strongest positive drivers of total cultivation cost.

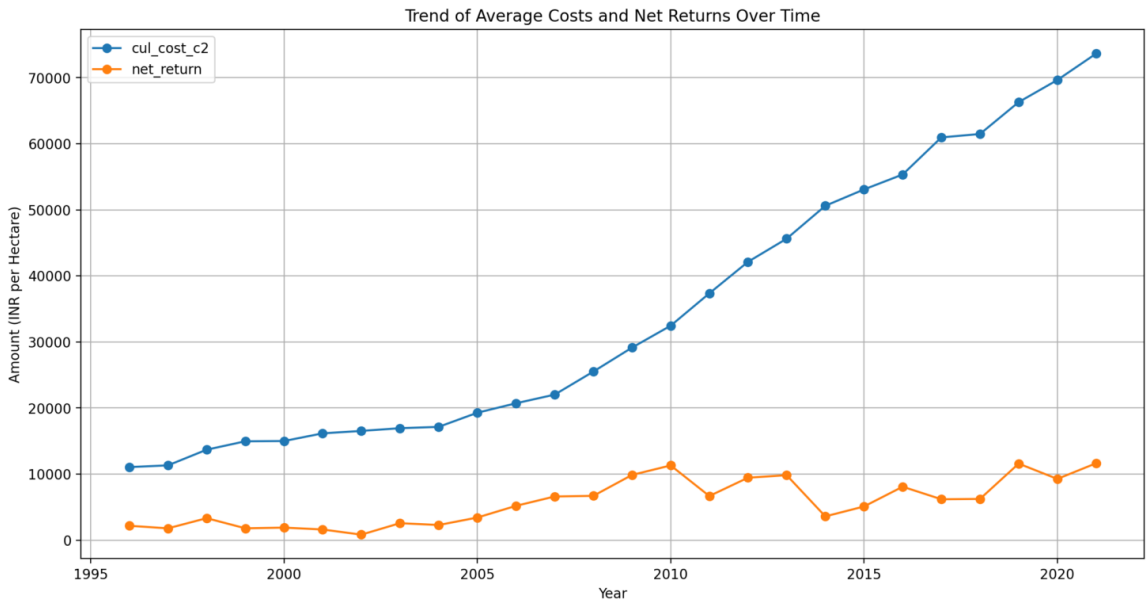


Figure 4.Temporal Analysis of Average Cost vs. Net Return (1995-2021)

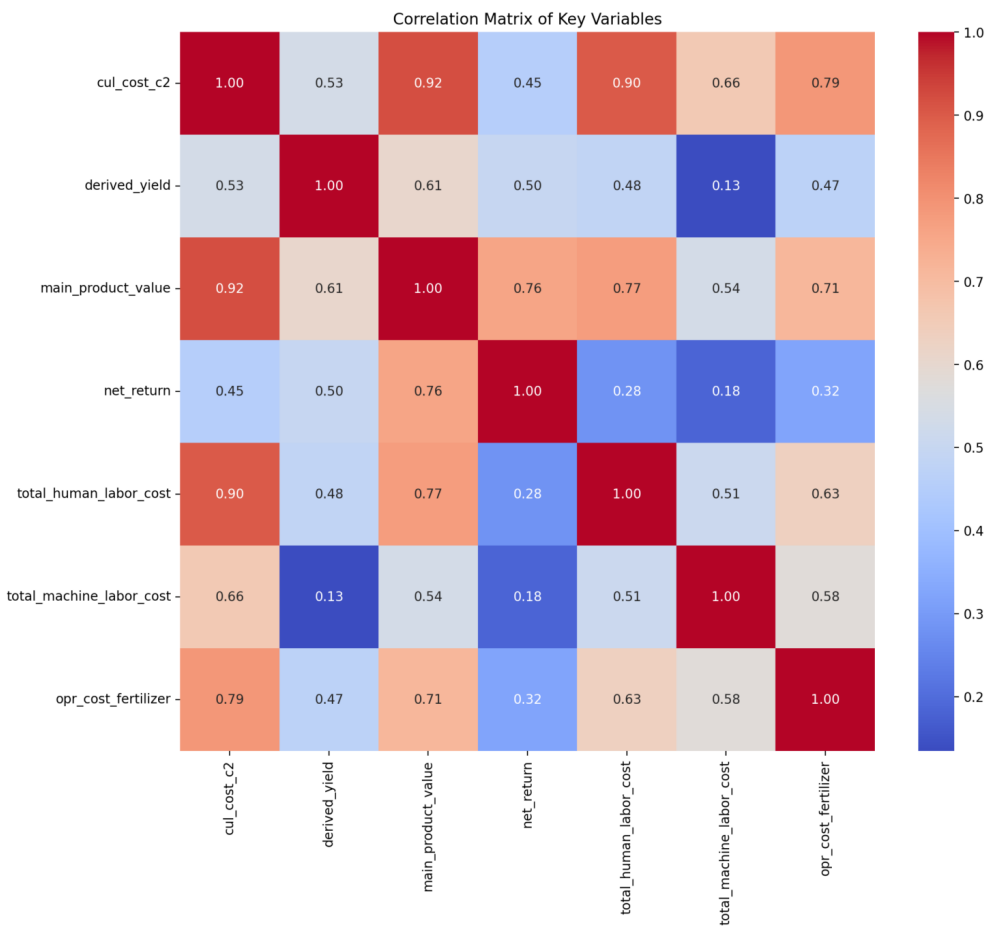


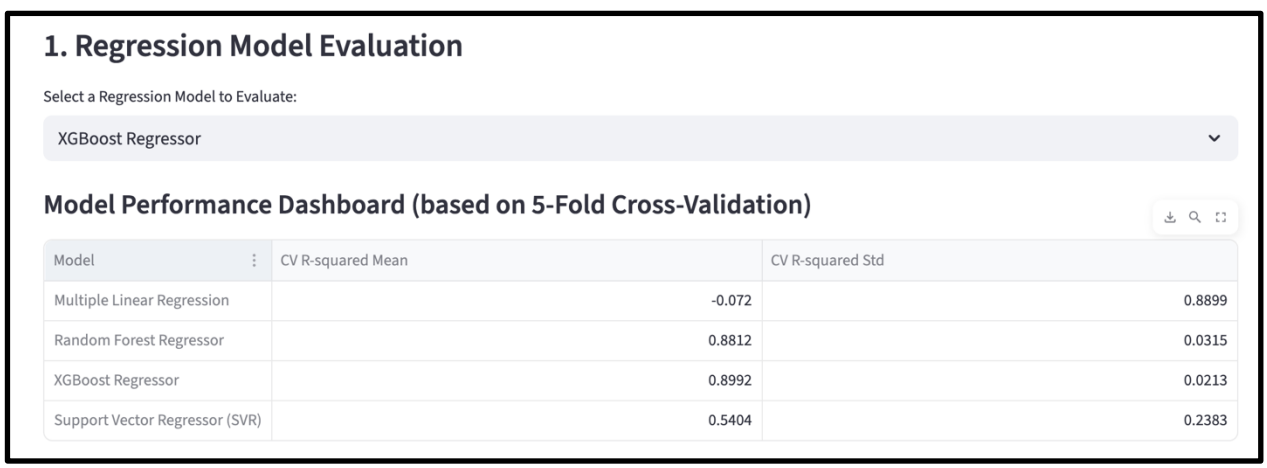
Figure 5.Correlation Matrix of Key EDA Variables

4. Modeling Methodology and Evaluation

A three-pronged modeling approach was adopted to gain a holistic understanding of the data. All models are implemented in models.py and served on the "Model Performance Evaluation" page. All performance dashboards are generated using a robust **5-Fold Cross-Validation** strategy for reliable metric comparison.

4.1. Task 1: Regression (Cost Prediction)

- **Objective:** To predict the prod_cost_c2 (production cost per quintal) using key operational and contextual features.
- **Models:** Multiple Linear Regression (Baseline), Random Forest Regressor, XGBoost Regressor (High Performance), Support Vector Regressor (Alternative Methodology).
- **Results:** The cross-validation dashboard (Fig 6) immediately identified the top performers. **XGBoost Regressor** (Mean CV R^2: **0.89**)and **Random Forest Regressor** (Mean CV R^2: **0.88**) are statistically tied for first place, far surpassing SVR and Linear Regression. This confirms that ensemble methods are best suited for this dataset.



1. Regression Model Evaluation

Select a Regression Model to Evaluate:

XGBoost Regressor

Model Performance Dashboard (based on 5-Fold Cross-Validation)

Model	CV R-squared Mean	CV R-squared Std
Multiple Linear Regression	-0.072	0.8899
Random Forest Regressor	0.8812	0.0315
XGBoost Regressor	0.8992	0.0213
Support Vector Regressor (SVR)	0.5404	0.2383

Figure 6.Regression Model Performance Dashboard (Cross-Validation)

- **Deep Dive (XGBoost):** The detailed evaluation of the best model, XGBoost, shows a high single-split **R-squared of 0.91** and an **RMSE of 703.33**. The Feature Importance plot (Fig 7) reveals that derived_yield is by far the most critical factor in determining production cost per quintal.

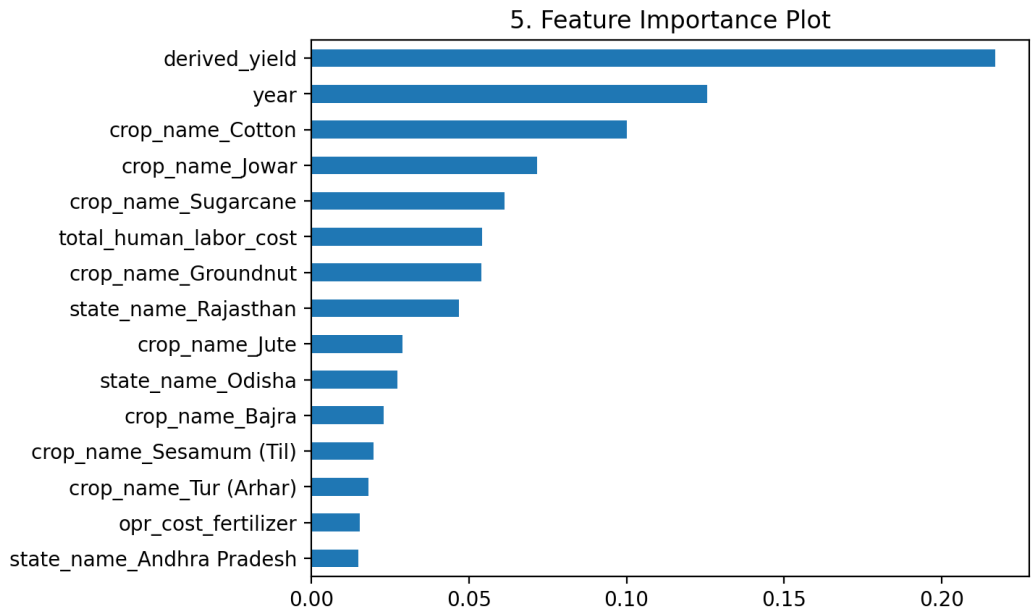
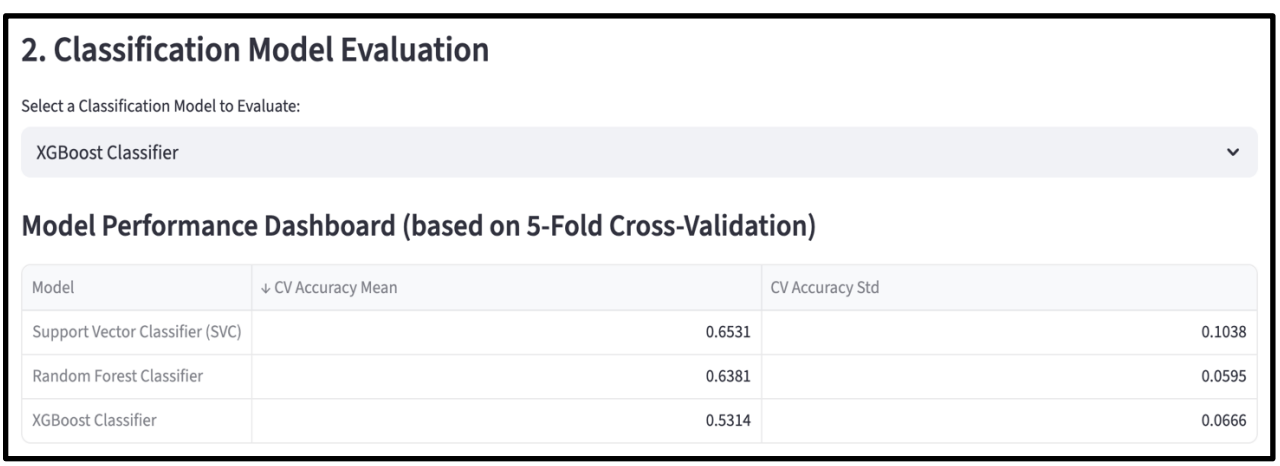


Figure 7.Feature Importance for XGBoost Regressor

4.2. Task 2: Classification (Efficiency Categorization)

- **Objective:** To predict which cost efficiency category a farm belongs to. The target variable, Cost_Efficiency_Class, was engineered by binning prod_cost_c2 into four quartiles (Lowest Cost, Below Average Cost, Above Average Cost, Highest Cost). This was done *within each crop type* to ensure a fair comparison.
- **Models:** Random Forest Classifier, XGBoost Classifier, Support Vector Classifier (SVC).
- **Results:** The classification dashboard (Fig 8) shows all three models are competitive. The **Support Vector Classifier (SVC)** has the highest Mean CV Accuracy (**0.65**) and a low standard deviation, making it the most reliable choice.



2. Classification Model Evaluation

Select a Classification Model to Evaluate:

XGBoost Classifier

Model Performance Dashboard (based on 5-Fold Cross-Validation)

Model	↓ CV Accuracy Mean	CV Accuracy Std
Support Vector Classifier (SVC)	0.6531	0.1038
Random Forest Classifier	0.6381	0.0595
XGBoost Classifier	0.5314	0.0666

Figure 8.Classification Model Performance Dashboard (Cross-Validation)

- **Deep Dive (SVC):** The detailed evaluation of the best model, SVC, shows an **Overall Accuracy** of **70.9%** and a **Matthews Correlation Coefficient (MCC)** of **0.612** indicating a very strong and balanced classification ability. The Normalized Confusion Matrix (Fig 9) shows the model is strongest at correctly identifying the Lowest Cost and Highest Cost categories.

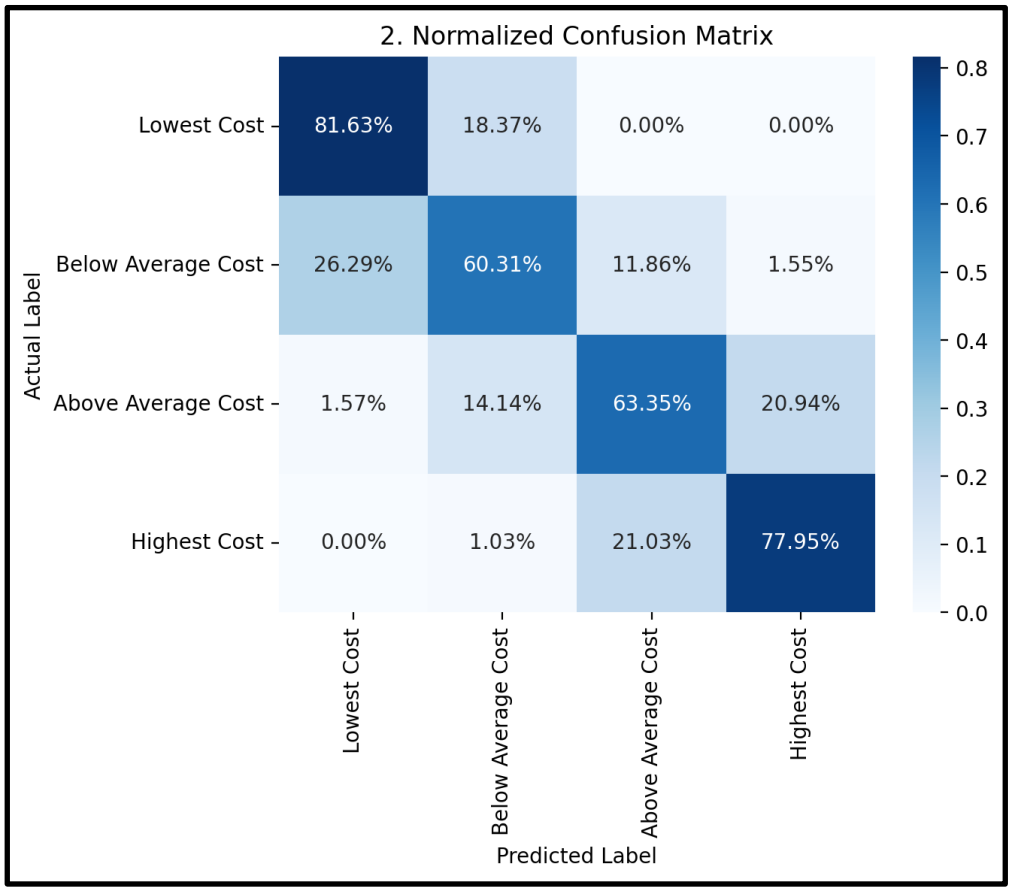


Figure 9.Confusion Matrix for SVC Model

4.3. Task 3: Clustering (Farm Archetype Discovery)

- **Objective:** To discover hidden "farm archetypes" using unsupervised learning. The features were specifically engineered to be ratios (e.g., mechanization_index, labor_cost_ratio), ensuring the clusters are based on *operational structure* rather than absolute size or cost.
- **Models:** K-Means, DBSCAN, Agglomerative Clustering.
- **Results:** K-Means clustering (with K=4) successfully partitioned the data into four distinct archetypes, achieving a **Silhouette Score of 0.34** The 2D PCA plot (Fig 10) provides a clear visualization of this separation. DBSCAN was effective at identifying outliers, while

Agglomerative Clustering produced a dendrogram showing the data's hierarchy.

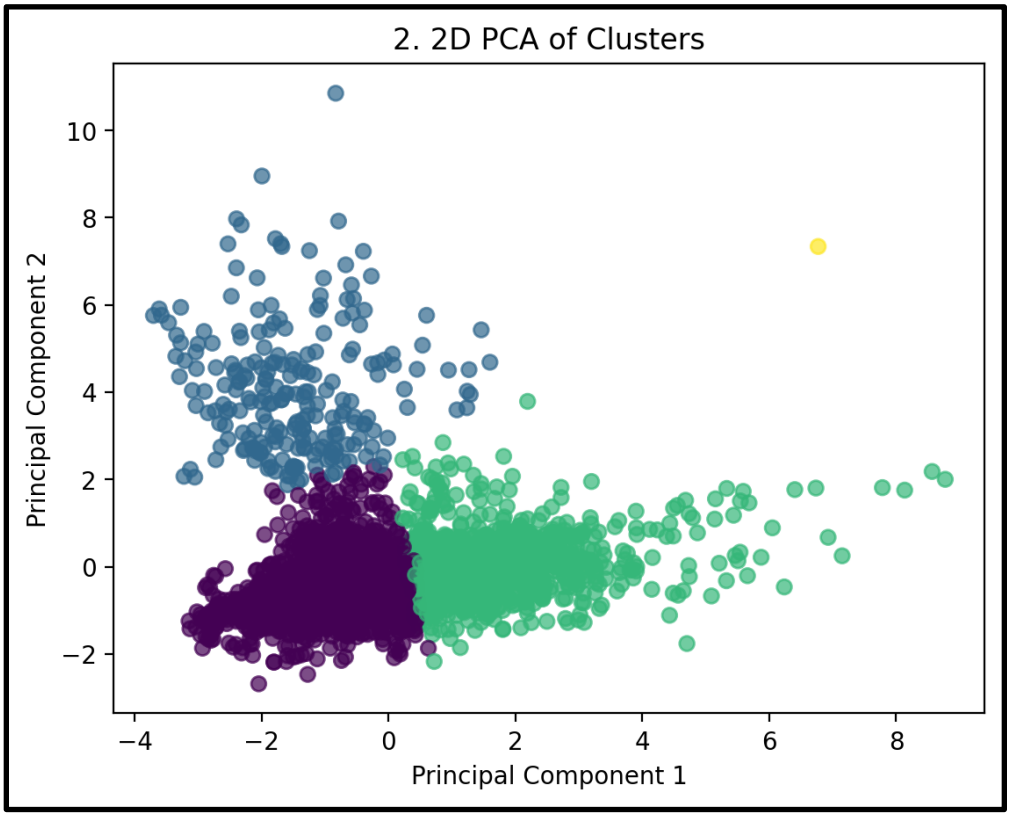


Figure 10.Clustering Analysis (K-Means PCA)

5. Interactive Dashboard Features (The Application)

The final models were integrated into three user-facing tools, built in app.py and models.py.

5.1. Predict Future Crop Cultivation Cost

This tool uses the best regression model (**Random Forest Regressor**) to forecast total cultivation cost for a user-selected state, crop, and future year. It applies a trend extrapolation

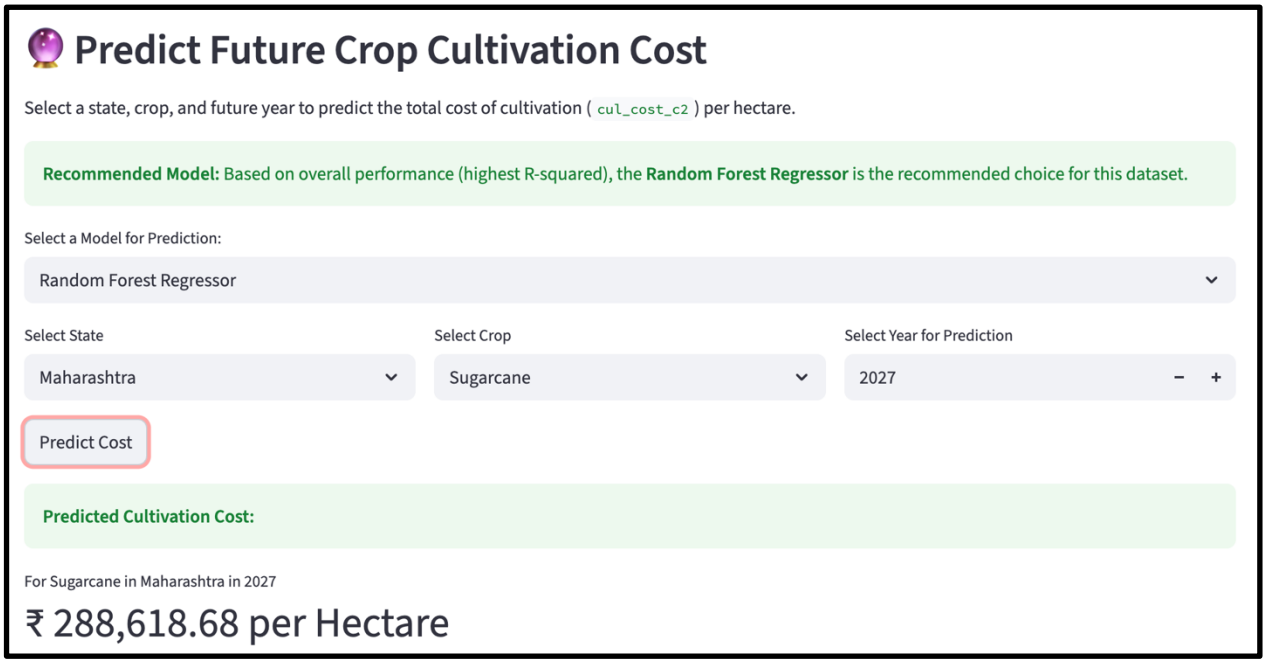


Figure 11.Future Cost Prediction Tool Interface

5.2. "What-If" Scenario Analysis

This tool allows users to perform sensitivity analysis. They can select a state and crop to load the latest cost data and then use sliders to simulate the impact of changes in input costs (e.g., "Fertilizer +20%," "Labor -10%"). The dashboard instantly recalculates the total cost.

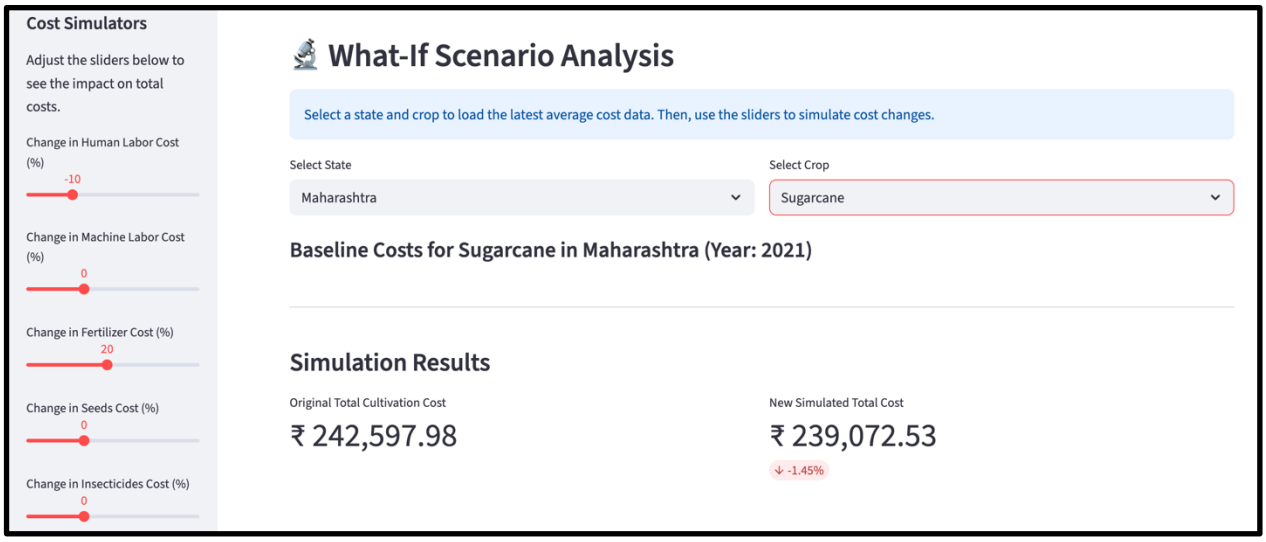


Figure 12.What-If" Scenario Analysis Tool

5.3. Profitability Calculator

This is the most advanced tool, combining prediction with user input.

1. **Step 1:** The app predicts the total cultivation cost for the *next* year.
2. **Step 2:** The user enters their *expected yield* and *expected market price*.
3. **Step 3:** The app calculates the final **Total Revenue**, **Predicted Net Profit/Loss**, and the **Break-Even Price** required to cover costs.

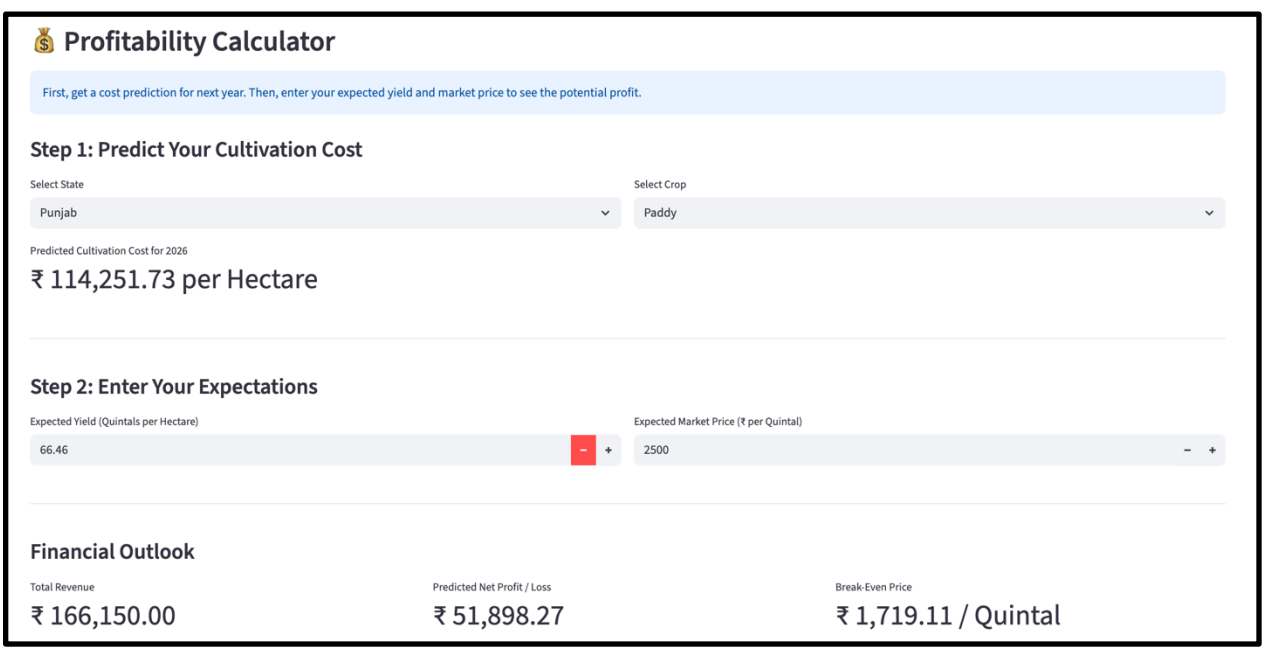


Figure 13.Profitability Calculator and Financial Outlook

6. Conclusion

This project successfully achieved all its objectives. We processed and analyzed a complex, real-world agricultural dataset and built a comprehensive, multi-page Streamlit dashboard.

The **Exploratory Data Analysis** provided critical insights, notably visualizing the "cost-price squeeze" that threatens farm profitability.

The **Modeling** phase was a comprehensive and robust comparison of 10 different machine learning models across 3 distinct tasks. We empirically determined that **XGBoost Regressor** is the most effective model for cost prediction, while **Support Vector Classifier (SVC)** is the most reliable for efficiency classification. The clustering analysis further revealed hidden archetypes in the data.

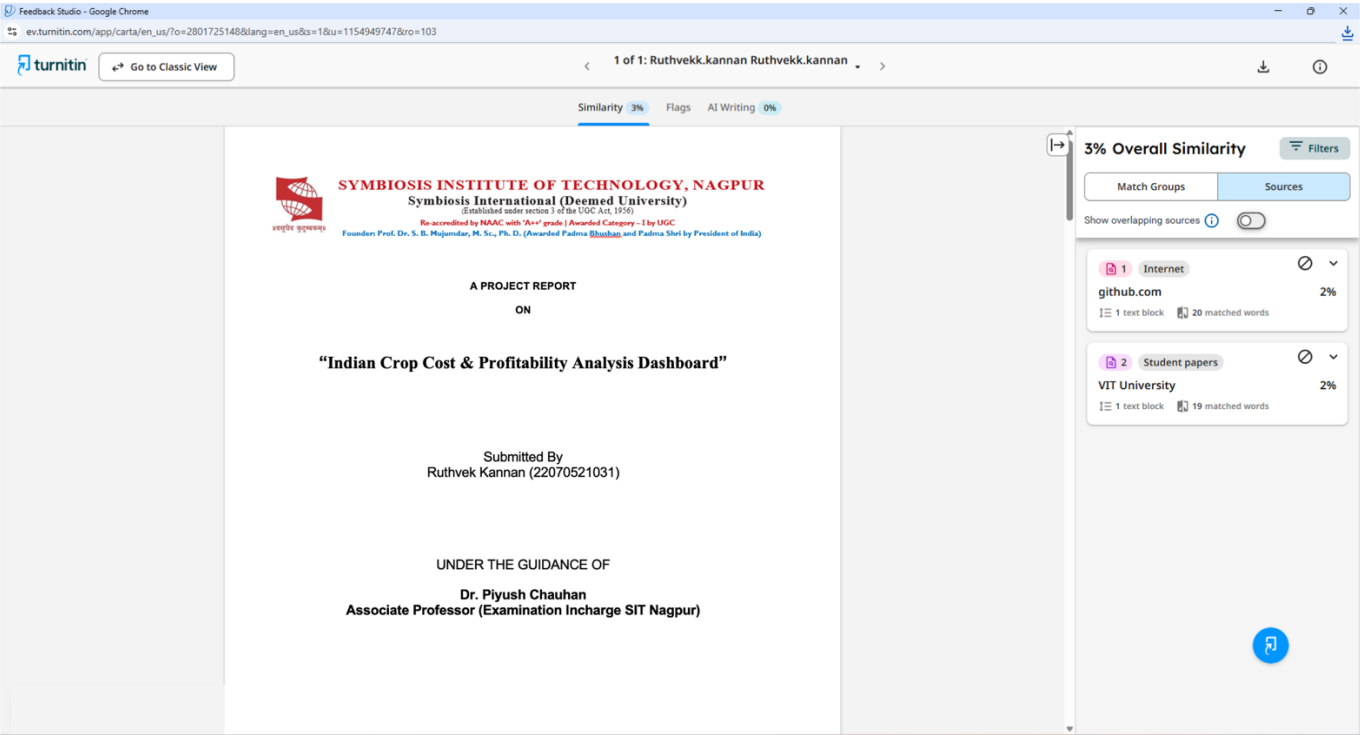
Finally, the project's primary objective was met by translating these complex models into three high-value, user-friendly interactive tools: a **Predictor**, a **"What-If" Simulator**, and a **Profitability Calculator**. These tools empower non-technical users, such as farmers or policymakers, to leverage advanced data science for data-driven planning, risk assessment, and decision-making.

7. Future Work

While this project provides a robust foundation, future iterations could expand its capabilities:

1. **Integrate More Data:** Incorporate external datasets, such as real-time Mandi market prices (via API), granular weather data, and soil type information to enhance model accuracy and the utility of the planning tools.
2. **Advanced Hyperparameter Tuning:** Implement automated hyperparameter tuning (e.g., GridSearchCV) to further optimize the performance of the top-performing models.
3. **Time-Series Forecasting:** For cost prediction, explore dedicated time-series models (like ARIMA, SARIMA, or Prophet) to capture seasonality and more complex temporal trends beyond the current linear extrapolation.
4. **Cloud Deployment:** Deploy the Streamlit application to a cloud service (e.g., Streamlit Community Cloud) to make it publicly accessible to a wider audience.
5. **Expand "What-If" Tool:** Allow users to simulate changes in yield and market price, not just cost, to create a more comprehensive sensitivity analysis.

Plagiarism & AI Report



Feedback Studio - Google Chrome

ev.turnitin.com/app/carta/en_us/?o=2801725148&lang=en_us&s=1&u=1154949747&ro=103

turnitin | Go to Classic View | 1 of 1: Ruthvekk.kannan Ruthvekk.kannan

Similarity 3% | Flags | AI Writing 0%

SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR
Symbiosis International (Deemed University)
(Established under section 3 of the UGC Act, 1956)
Re-accredited by NAAC with 'A++' grade | Awarded Category – I by UGC
Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

A PROJECT REPORT

ON

"Indian Crop Cost & Profitability Analysis Dashboard"

Submitted By
Ruthvek Kannan (22070521031)

UNDER THE GUIDANCE OF

Dr. Piyush Chauhan
Associate Professor (Examination Incharge SIT Nagpur)

3% Overall Similarity | Filters

Match Groups | Sources

Show overlapping sources

1 Internet
github.com 2%
1 text block 20 matched words

2 Student papers
VIT University 2%
1 text block 19 matched words