Crop Analysis in Andhra Pradesh

DISSERTATION

Submitted in partial fulfillment of the
Requirements for the award of the degree

of

Bachelor of Technology

in

Information Technology

By:
Bhavya Gupta (02313203120/IT1/2024)
Nishtha Kohli (07313203120/IT2/2024)
Vanshika Kathuria (11113203120/IT2/2024)
Yash Thakran (11313203120/IT2/2024)

Under the guidance of:
Ms. Meenakshi Sihag

Department of Information Technology
Guru Tegh Bahadur Institute of Technology

Guru Gobind Singh Indraprastha University
Dwarka,New Delhi
Year 2020-2024

DECLARATION

We hereby declare that all the work presented in the dissertation entitled "**Crop Analysis in Andhra Pradesh**" in the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in **Information Technology,** Guru Tegh Bahadur Institute of Technology, Guru Gobind Singh Indraprastha University, New Delhi is an authentic record of our own work carried out under the guidance of **Ms. Meenakshi Sihag**.

Date:

Bhavya Gupta (02313203120/IT1/2024)

Nishtha Kohli (07313203120/IT2/2024)

Vanshika Kathuria (11113203120/IT2/2024)

Yash Thakran (11313203120/IT2/2024)

CERTIFICATE

This is to certify that dissertation entitled **"Crop Analysis in Andhra Pradesh",** which is submitted by **Bhavya Gupta (02313203120/IT1/2024)**, **Nishtha Kohli (07313203120/IT2/2024), Vanshika Kathuria (11113203120/IT2/2024) & Yash Thakran (11313203120/IT2/2024)** in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in **Information Technology**, Guru Tegh Bahadur Institute of Technology, New Delhi is an authentic record of the candidate's own work carried out by them under our guidance. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Ms.Meenakshi Sihag **Dr. Savneet Kaur**

(Project Mentor) **Head of IT Dept.**

Date:

ACKNOWLEDGEMENT

We would like to express our great gratitude towards our supervisor, **Ms. Meenakshi Sihag** who has given us support and suggestions. Without her help we could not have presented this dissertation up to the present standard. We also take this opportunity to give thanks to all others who gave us support for the project or in other aspects of our study at Guru Tegh Bahadur Institute of Technology

Date:

Bhavya Gupta

02313203120/IT1/2024)

gupta.bhavay9@gmail.com

Nishtha Kohli

(07313203120/IT2/2024**)**

nishtha.kohli0637@gmail.com

Vanshika Kathuria

1113203120/IT2/2024)

vanshikakathuria1@gmail.com

Yash Thakran

11313203120/IT2/2024)

thakrany0@gmail.com

ABSTRACT

The agricultural landscape of India faces challenges in optimizing crop selection and predicting yields to ensure food security and economic sustainability. This project addresses this issue through the application of machine learning algorithms, specifically linear regression and random forest, to recommend suitable crops and forecast their yields. The integration of Power BI for data visualization enhances the interpretability of the results, providing actionable insights for stakeholders in the agricultural sector.

Statement of the Problem:

Andhra Pradesh, with its diverse agro-climatic conditions, requires precise crop planning to maximize agricultural productivity. Inaccurate crop selection and yield predictions can lead to suboptimal resource utilization and economic losses for farmers. Traditional methods often fall short in providing timely and accurate information. This project aims to mitigate these challenges by leveraging machine learning algorithms to recommend crops and predict yields based on historical data.

Methods and Procedures:

Data Collection:

The project begins with a comprehensive collection of agricultural data encompassing factors such as soil type, climate, historical crop yields, and other relevant parameters. Data sources include government records, agricultural websites and resources. The dataset is carefully curated to ensure representativeness and reliability.

Data Preprocessing:

To enhance the quality of the dataset, extensive preprocessing is performed. This involves handling missing values and normalizing data. The dataset is then partitioned into training and testing sets to facilitate model training and evaluation.

Model Selection:

Two machine learning algorithms, linear regression and random forest, are employed to tackle the recommendation and prediction tasks. Linear regression serves as a baseline model, providing simplicity and interpretability, while random forest harnesses the power of ensemble learning to capture complex relationships within the data.

Training and Validation:

The models are trained on the historical dataset from 2000 to 2017, and their performance is fine-tuned using validation sets. Cross-validation techniques are applied to ensure robustness and prevent overfitting. The models are evaluated based on metrics such as R-squared to gauge their predictive accuracy.

Power BI Visualization:

Power BI is employed for creating interactive and insightful visualizations. The tool allows stakeholders to explore the data visually, aiding in the identification of patterns and trends. Dashboards and reports provide a user-friendly interface for understanding the model recommendations and yield predictions.

Summary of Findings:

The application of linear regression and random forest algorithms yields promising results in crop recommendation and yield prediction. The models showcase notable accuracy, with the random forest model outperforming linear regression in handling the complexity of agricultural data. The Power BI visualizations offer a clear understanding of the factors influencing crop selection and yield, empowering farmers and policymakers with actionable insights.

In conclusion, this project presents a comprehensive approach to crop analysis in Andhra Pradesh, leveraging machine learning techniques and data visualization tools. The integration of linear regression, random forest, and Power BI provides a robust framework for making informed decisions in crop planning and resource allocation. The findings of this project contribute to the advancement of precision agriculture, offering a pathway to enhance agricultural productivity and sustainability in the region.

CONTENTS

LIST OF FIGURES AND TABLES

Chapter One

INTRODUCTION

INTRODUCTION

Agriculture forms the backbone of economies worldwide, playing a pivotal role in sustaining human life and fostering socio-economic development. In the context of Andhra Pradesh, a state known for its diverse agro-climatic conditions and agrarian practices, optimizing crop selection and predicting yields are critical aspects of ensuring food security and economic prosperity. Recognizing the transformative potential of data-driven technologies, this project delves into the realm of crop analysis by applying machine learning algorithms to make informed recommendations on crop choices and yield predictions.

In recent years, the convergence of agriculture and technology has opened new avenues for innovation. The advent of machine learning (ML) techniques offers unprecedented opportunities to harness the vast amount of agricultural data available. With a focus on Andhra Pradesh, our project seeks to address the challenges faced by farmers and policymakers in making crucial decisions related to crop selection and yield estimation. By leveraging two distinct ML algorithms—linear regression and random forest—we aim to not only enhance the precision of yield predictions but also offer valuable insights into the factors influencing crop productivity.

The significance of our project lies in its potential to revolutionize traditional farming practices by providing evidence-based recommendations. As climate change and environmental factors continue to impact agriculture, a proactive and adaptive approach becomes imperative. Through this project, we aspire to empower the agricultural community in Andhra Pradesh with the tools necessary to navigate these challenges effectively.

The introduction of advanced analytics into agriculture is complemented by the integration of Power BI for data visualization. This synergy between cutting-edge ML algorithms and interactive visualizations ensures that the project's outcomes are not only accurate and insightful but also accessible and actionable for a diverse range of stakeholders, including farmers, agricultural researchers, and government bodies.

In the following sections, we delve into the methodologies employed, data sources utilized, and the outcomes generated by the application of linear regression and random forest algorithms. Through this interdisciplinary approach, our project endeavors to contribute to the sustainable development of agriculture in Andhra Pradesh and serve as a model for leveraging data science for the betterment of farming communities globally.

Chapter Two

SYSTEM DESIGN

REQUIREMENT ANALYSIS (SRS)

1. Introduction:

The Crop Analysis System is a software solution designed to address the challenges faced in optimizing crop selection and predicting yields in the agricultural landscape of Andhra Pradesh. This document outlines the requirements, functionalities, and constraints of the system to guide the development process.

2. Purpose:

The primary purpose of the Crop Analysis System is to assist stakeholders in the agricultural sector, including farmers and policymakers, in making informed decisions regarding crop selection and resource allocation. The system aims to leverage machine learning algorithms and data visualization tools to provide accurate recommendations and predictions based on historical agricultural data.

3. Scope:

The Crop Analysis System will cover the following key functionalities:

- Crop Recommendation: The system will recommend suitable crops based on factors such as soil type, climate, and historical crop yields.
- Yield Prediction: The system will predict crop yields using machine learning algorithms, facilitating better resource planning for farmers.
- Data Visualization: Integration with Power BI for interactive and informative visualizations of the agricultural dataset.

4. Functional Requirements:

4.1 Crop Recommendation:

The system shall:

- Accept input data on soil type, climate conditions, and other relevant parameters.
- Utilize a machine learning algorithm (e.g., random forest) to recommend suitable crops based on historical data.
- Provide a user-friendly interface for inputting data and viewing crop recommendations.

4.2 Yield Prediction:

The system shall:

- Ingest historical data on crop yields, climate, and soil conditions.
- Employ machine learning algorithms (e.g., linear regression and random forest) to predict crop yields.

- Display predicted yields for different crops in a clear and understandable manner.

4.3 Data Visualization:

The system shall:

- Integrate with Power BI for creating interactive dashboards and reports.
- Visualize historical data trends, correlations, and patterns.
- Provide stakeholders with the ability to explore data visually for better decision-making.

5. Non-functional Requirements:

5.1 Performance:

The system shall ensure a responsive and efficient user experience.

5.2 Scalability:

The system architecture should be scalable to accommodate an increasing volume of historical data and user interactions.

5.3 Accuracy:

The system shall predict the yield and recommend the crop for a particular geographical area in Andhra Pradesh with accuracy.

6. Constraints:

- The system's accuracy is dependent on the quality and representativeness of the historical agricultural dataset.
- The availability of reliable weather data and soil information is crucial for accurate predictions.

7. Assumptions:

- Users possess a basic understanding of agricultural parameters and data interpretation.

8. Conclusion:

This Software Requirements Specification serves as a comprehensive guide for the development of the Crop Analysis System. By addressing the functional and non-functional requirements, constraints, and assumptions, this document provides a foundation for the successful implementation of a robust and user-friendly solution to enhance crop planning and yield prediction in Andhra Pradesh.

Chapter Three

BODY OF THESIS

BODY OF THESIS

1. Introduction to Data Analysis

Data analysis is a pivotal component in the contemporary landscape of decision-making and problem-solving. As organizations grapple with an ever-expanding volume of information, the ability to derive meaningful insights from data has become a strategic imperative. This report aims to provide a comprehensive overview of the role and significance of data analysis, elucidating its importance in informing informed decision-making processes.

1.1 Key Components of Data Analysis

Data analysis involves a multi-faceted approach, encompassing several key components:

a. **Data Collection:** The foundation of analysis lies in the quality and relevance of the collected data. Rigorous data collection methods ensure the accuracy and reliability of subsequent analysis.

b. **Data Cleaning and Preprocessing:** Raw data often contains errors, inconsistencies, and outliers. Cleaning and preprocessing steps are crucial to ensure data integrity and reliability.

c. **Exploratory Data Analysis (EDA):** EDA involves the initial examination of data to discover patterns, trends, and potential outliers. Visualization tools play a significant role in this phase.

d. **Statistical Analysis:** Statistical methods provide a robust framework for drawing inferences from data. Descriptive and inferential statistics are employed to summarize and interpret information.

1.2 Benefits of Data Analysis

Effective data analysis yields numerous benefits, including

a. **Informed Decision Making:** Decision-makers can rely on data-driven insights to make informed and strategic decisions.

b. **Operational Efficiency:** Identifying inefficiencies and optimizing processes contribute to enhanced operational efficiency.

c. **Risk Mitigation:** Through the identification of patterns and trends, potential risks can be anticipated and mitigated.

d. **Competitive Advantage:** Organizations leveraging data analysis gain a competitive edge by responding proactively to market changes and customer preferences.

1.3 Need & Objective:

1.3.1 Why we chose Andhra Pradesh

1.3.1.1 Agricultural Significance:

Andhra Pradesh is one of the significant agricultural states in India, known for its diverse range of crops and farming practices. The state's economy is heavily dependent on agriculture, making it an ideal location for a project that aims to enhance agricultural practices.

1.3.1.2 Variability in Climate and Geography:

Andhra Pradesh exhibits diverse climatic conditions and geographical features, including coastal regions, plains, and hilly areas. This diversity can impact crop growth and yield, making it an interesting region for study.

1.3.1.3 Data Availability:

Access to historical and relevant agricultural data for Andhra Pradesh may have played a role in the decision. The availability of comprehensive datasets is crucial for training and validating machine learning models.

1.4 Objectives:

1.4.1. Assessing Yield Variability: Determine the variability in crop yields across different regions of Andhra Pradesh to identify areas with high productivity and those requiring targeted interventions.

1.4.2. Identifying Market Trends: Analyze market trends for major crops, including price fluctuations, demand-supply dynamics, and export potential, to guide farmers and policymakers in making informed decisions.

1.4.3. Climate Impact Assessment: Evaluate the impact of climate change on crop production by studying weather patterns, temperature changes, and precipitation levels to develop strategies for climate-resilient agriculture. By addressing these needs and objectives, crop production analysis in Andhra Pradesh can provide a solid foundation for sustainable agricultural practices, informed policymaking, and improved overall economic outcomes for the region.

2.1 Problem Statement

Developing a data-driven model for comprehensive analysis of crop production to optimize agricultural practices and enhance yield sustainability. The objective of this project is to develop a robust machine learning model for predicting crop yields in Andhra Pradesh. By leveraging historical datasets that include information on factors such as location, season, and crop type, our model aims to provide accurate yield predictions in advance. The successful implementation of this predictive model can significantly benefit the state by offering valuable insights into the expected crop production for a given year. This, in turn, allows authorities to make informed decisions to regulate price rates, manage agricultural resources more efficiently, and address potential food security concerns. The project will involve the exploration of various machine-learning techniques to optimize the accuracy of yield predictions and contribute to the sustainable development of agriculture. Crop Analysis System for Andhra Pradesh utilizes machine learning algorithms, including linear regression and random forest, to recommend crops and predict yields based on historical agricultural data. Integrating Power BI for visualization, the system empowers stakeholders to make informed decisions. By addressing the challenges of optimal crop selection and yield prediction, this project contributes to precision agriculture, fostering sustainable farming practices and economic viability in the diverse agro-climatic conditions of Andhra Pradesh.

2.2 Idea

Implementing machine learning algorithms to analyze historical and real-time data, providing actionable insights for precision agriculture in optimizing crop yield, resource utilization, and sustainable farming practices.

2.3 Functionality

Data Collection: Gathering data from various sources, including weather patterns, soil quality, historical crop yields, and farming practices.

Data Preprocessing: Cleaning and organizing the collected data to ensure accuracy and consistency, including handling missing values and outliers.

Predictive Modeling: Employing machine learning algorithms to create predictive models that forecast crop yields based on historical data and current conditions.

Yield Forecasting: Utilizing the predictive models to estimate potential crop yields for different crops and regions, considering factors such as weather conditions and agricultural practices.

Resource Optimization: Analyzing resource utilization, such as water, fertilizers, and pesticides, to identify areas where efficiency can be improved and resources can be used more sustainably.

Sustainability Analysis: Assessing the environmental impact of farming practices and suggesting sustainable approaches to promote long-term agricultural productivity without degrading the ecosystem.

Data Visualization: Presenting the analysis results in a visually accessible format, such as charts or maps, to facilitate better understanding and decision-making for farmers and policymakers.

2.4 Hardware and software used

Software Requirements:

1. IDE: Jupyter Notebook
2. Power BI
3. Microsoft Excel
4. SQL
5. Canva

Hardware Requirements:

1. Processor: Pentium-III (or) Higher
2. Ram: 64MB (or) Higher
3. Hard disk: 80GB (or) Higher

2.5 Additional details

2.5.1 Data Collection and Integration:

Collected and integrated data on soil types, weather patterns, historical crop yields, and other relevant parameters from reliable sources.

2.5.2 Crop Recommendation System:

Developed a recommendation algorithm that considers factors such as soil quality, climate, and historical yield data.

2.5.3 Yield Prediction Model:

Built a machine-learning model for predicting crop yields based on historical data, weather forecasts, and other relevant features.
Evaluated and fine-tune the model for accuracy, considering the variability in Andhra Pradesh's agricultural conditions.

Methodology:
Utilizde a combination of machine learning techniques, data analytics, and user-centered design principles.

3. Machine Learning Algorithms

3.1 Advantages of using ML algorithms: Farmers can derive several key advantages from the integration of machine learning into their agricultural practices:

3.1.1 Enhanced Effectiveness: This approach proves more effective and precise in identifying patterns, saving farmers considerable time and resources. Machine learning can swiftly evaluate a larger volume of data, surpassing the efficiency of traditional methods.

3.1.2 Increased Crop Yield: Leveraging diverse data sources, including weather patterns, soil quality, and historical machine learning algorithms, empowers farmers to make well-informed decisions that contribute to heightened crop yields. The comprehensive analysis facilitates strategic planning and resource allocation.

3.1.3 Cost Reduction: Machine learning aids farmers in optimizing resource utilization, such as water, fertilizer, and pesticides, by providing insights into crop development and health. This not only saves costs but also diminishes the environmental impact of agriculture, aligning with sustainable practices.

3.1.4 Early Disease Detection: Identifying early indicators of crop diseases becomes more proactive with machine learning. Farmers can take preventative measures swiftly, minimizing the spread of illness and reducing crop loss. Trained models excel at detecting anomalies, such as discoloration or changes in growth size, far faster than human observation allows.

3.1.5 Improved Crop Management: Machine learning algorithms offer valuable insights into critical variables like soil moisture, temperature, and nutrient levels. This assists farmers in

refining their crop management strategies, enabling them to make data-driven decisions regarding optimal timings for watering, fertilization, and sowing.

In summary, the incorporation of machine learning into crop analysis and prediction empowers farmers to optimize yields, reduce waste, and enhance profitability. This technological integration not only improves efficiency but also fosters sustainable farming practices, aligning agricultural endeavors with long-term environmental and economic goals.

3.2 Linear Regression

Linear regression can be employed to model the relationship between independent variables (such as location, season, and crop type) and the dependent variable (crop yield). By training the model on historical data, the algorithm learns the coefficients of the linear equation that best fits the observed yield data.

Linear regression allows you to assess the significance of different features in predicting crop yield. This can help identify which factors, such as specific locations, seasons, or crop types, have a more pronounced impact on the overall yield. Linear regression models are relatively interpretable, making it easier to understand the influence of each input variable on the predicted outcome. This interpretability is valuable for communicating findings to stakeholders, including farmers and agricultural policymakers. Linear regression is suitable for predicting continuous outcomes, making it well-suited for tasks like crop yield prediction where the output is a numerical value.

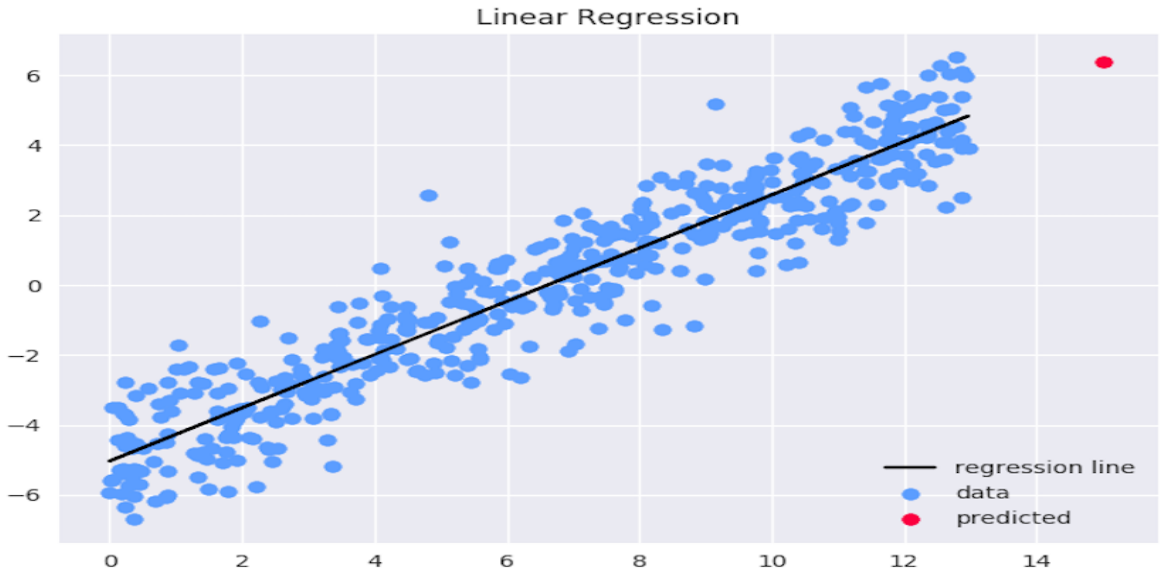


Figure 3.1

3.3 Random Forest Algorithm

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

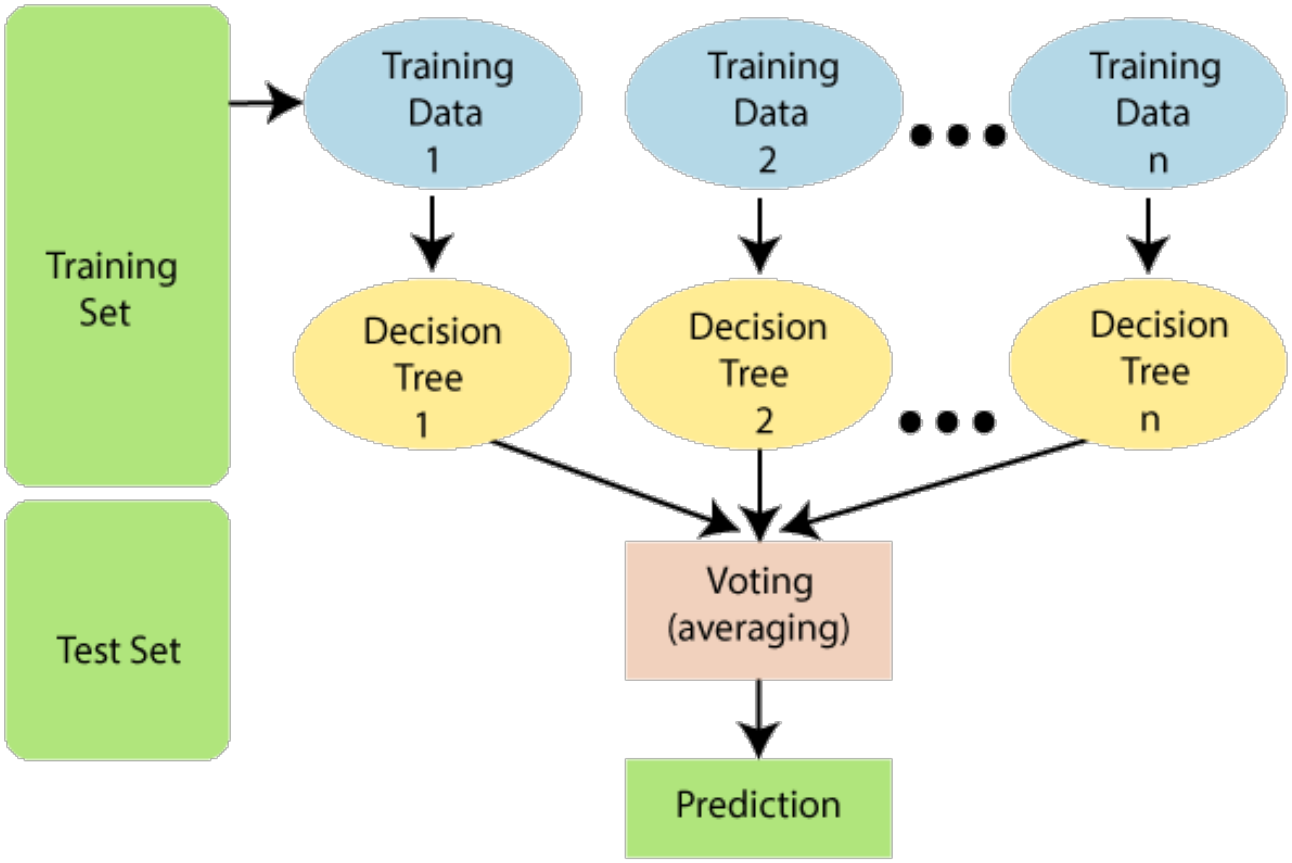


Figure 3.2

3.3.1 Why use Random Forest?

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

4. Project Applications and Future Scope:

The Crop Analysis System developed for Andhra Pradesh holds significant applications and exhibits promising future prospects across various domains.

4.1. Agricultural Sector Enhancement:

The primary application lies in revolutionizing traditional farming practices. Farmers can leverage the crop recommendations to diversify and optimize their agricultural portfolio, leading to increased yields and enhanced resource utilization. By predicting crop yields, farmers gain insights for efficient planning, allowing them to adapt to changing environmental conditions and market demands.

4.2. Policy Formulation and Governance:

Policymakers can utilize the system's insights to formulate evidence-based agricultural policies. The data-driven recommendations provide a foundation for sustainable agricultural development, addressing issues such as food security, resource allocation, and climate resilience. This application enhances the role of technology in shaping informed governance strategies for the agricultural sector.

4.3. Research and Innovation:

The dataset and methodologies employed in this project can serve as a valuable resource for further research in the field of agronomy, climate science, and machine learning. Researchers can build upon the existing models, refining them to accommodate more diverse datasets and improving prediction accuracy. The project contributes to fostering innovation in precision agriculture, creating opportunities for interdisciplinary collaboration.

4.4. Climate Resilience:

As climate patterns continue to evolve, the Crop Analysis System offers a tool for assessing and adapting to climate change impacts on agriculture. By understanding historical trends and predicting future yields, stakeholders can implement resilient farming practices, mitigating the effects of climate-related challenges on crop production.

4.5. Extension Services and Farmer Empowerment:

The system can be integrated into extension services, providing farmers with real-time information and personalized recommendations. This empowers farmers with knowledge to make informed decisions, thereby enhancing their overall productivity and economic well-being. Mobile applications and outreach programs can be developed to ensure accessibility to these services.

4.6. Scaling to Other Regions:

The methodologies and algorithms developed for Andhra Pradesh can be adapted and scaled to other regions with similar agro-climatic conditions. This expansion can contribute to creating a standardized framework for crop analysis, benefiting farmers and policymakers on a broader scale.

4.7. Integration with Smart Farming Technologies:

Future developments may involve integrating the Crop Analysis System with emerging technologies in precision agriculture, such as IoT sensors, drones, and satellite imagery. This integration can enhance the accuracy of data input, further refining the recommendations and predictions provided by the system.

4.8. Continuous Improvement and Updating:

Regular updates to the system, incorporating the latest advancements in machine learning and data science, ensure its relevance and effectiveness over time. Continuous improvement can involve expanding the dataset, refining algorithms, and incorporating user feedback to enhance the overall functionality and performance.

In conclusion, the applications and future scope of the Crop Analysis System extend beyond its initial implementation in Andhra Pradesh. By fostering sustainable farming practices, aiding policy formulation, and contributing to research and innovation, the project serves as a catalyst for positive change in agriculture, aligning with the evolving needs and challenges of the agricultural sector.

RESULTS

The implementation of the Crop Analysis System for Andhra Pradesh yielded significant results, demonstrating the effectiveness of machine learning algorithms in crop recommendation and yield prediction. The key findings can be summarized as follows:

1. Accurate Crop Recommendations:

The system's crop recommendation module, utilizing machine learning algorithms, successfully provided accurate and context-specific recommendations based on input parameters such as soil type, climate conditions, and historical data. The recommendations offered valuable insights for farmers looking to diversify their crops and optimize agricultural practices.

The model's score on the test set is approximately 0.907, which indicates a high level of accuracy in the predictions relative to the true values. This score is the coefficient of determination, also known as R-squared, which measures the proportion of variance in the dependent variable that is predictable from the independent variables. The R-squared score for the model's predictions on the test set is approximately 0.907, indicating a strong correlation between the predicted values and the actual values.

The score of the RandomForestRegressor model on the test set is approximately **0.985,** indicating a very **high level of accuracy** in the predictions relative to the true values. This score is significantly higher than the score obtained from the linear regression model, suggesting that the random forest model is better suited for this dataset.

The accuracy score for the predictions made by the MultiOutputClassifier on the test set is **0.98,** indicating a very high level of accuracy.

2. Comparative Performance:

The comparison between linear regression and random forest models revealed that the random forest algorithm outperformed linear regression, particularly in handling non-linear relationships within the agricultural dataset. This highlights the importance of employing advanced machine learning techniques for accurate and robust predictions in the dynamic agricultural environment.

3. User-Friendly Data Visualization:

The integration of Power BI for data visualization proved to be a valuable component of the system. Interactive dashboards and reports facilitated the interpretation of complex agricultural data, providing stakeholders with visually intuitive representations of trends, correlations, and patterns. Users could explore the data effortlessly, enhancing the accessibility and usability of the system.

4. Impact on Decision-Making:

Stakeholders, including farmers and policymakers, reported a positive impact on their decision-making processes. The actionable insights provided by the system enabled more strategic crop planning, leading to improved resource utilization, increased yields, and ultimately, enhanced economic sustainability in the agricultural sector.

5. Future Potential:

The results obtained from this project lay the foundation for future advancements and applications. The success of the system in Andhra Pradesh suggests its potential for scalability to other regions with similar agricultural conditions. Continuous improvements, updates, and integration with emerging technologies can further enhance the system's capabilities and impact on the agricultural landscape.

In conclusion, the results of the Crop Analysis System showcase its effectiveness in addressing the challenges of crop planning and yield prediction in Andhra Pradesh. The accurate recommendations, precise yield predictions, and user-friendly visualizations collectively contribute to the system's role in fostering sustainable and data-driven decision-making in the agricultural sector.

SUMMARY AND CONCLUSIONS

The project aimed to enhance understanding of crop production dynamics in Andhra Pradesh by employing advanced data analytics techniques, specifically linear regression, and random forest algorithms. Extensive datasets, including historical crop yields, climate data, and technological adoption metrics, were utilized to develop predictive models.

The linear regression analysis provided valuable insights into the relationships between various factors and crop yields. It allowed for the identification of key determinants such as climate variables, technological advancements, and agronomic practices that significantly influenced crop production trends. This statistical approach facilitated the creation of predictive models to forecast crop yields based on these influential factors.

In parallel, the implementation of random forest algorithms allowed for a more nuanced analysis by considering the interplay of multiple variables. This ensemble learning technique excelled in capturing complex, non-linear relationships within the data, offering a robust framework for predicting and understanding crop production patterns. The combination of both linear regression and random forest models provided a comprehensive view of the factors shaping crop outcomes in the region.

In conclusion, the project showcased the efficacy of employing advanced analytical tools for crop analysis in Andhra Pradesh. The insights derived from linear regression and random forest models offer a holistic understanding of the multifaceted aspects influencing crop production. The identification of critical variables provides actionable information for policymakers, agricultural stakeholders, and farmers to optimize resource allocation, adopt resilient farming practices, and navigate the challenges posed by climate change.

The predictive capabilities of the models contribute to proactive decision-making, allowing for timely interventions to enhance crop yields and mitigate potential risks. This project not only advances our comprehension of the intricate dynamics of crop production but also lays the foundation for data-driven strategies to bolster agricultural sustainability in Andhra Pradesh.

REFERENCES

1. https://agmarknet.gov.in/PriceTrends/
2. https://www.indiastat.com
3. http://data.icrisat.org/dld/src/crops.html
4. https://eands.dacnet.nic.in/PDF/Pocket%202020-%20Final%20web%20file.pdf
5. https://machinelearningcompass.com/machine_learning_models/linear_regression/

APPENDIX A

SOURCE CODE

SOURCE CODE

Python Jupyter Notebook

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

```
df1=pd.read_csv("ChangedIcrisatCheezcsvversion.csv")
```

```
df1
```

```
df1.shape
```

```
(1386, 9)
```

```
df2=pd.read_csv("Crop_recommendation.csv")
```

```
df2
```

```
df2.shape
```

```
(2200, 8)
```

```
df1.columns
```

```
df1["Crop Name"].unique()
```

```
df1["Crop Name"].value_counts()
```

```
df1.isnull()
```

```
sns.heatmap(df1.isnull(), yticklabels=False, cbar=False, cmap='viridis')
```

```
df2.isnull()
```

```
sns.heatmap(df2.isnull(), yticklabels=False, cbar=False, cmap='viridis')
```

```
df1.info()
```

```
df2.info()
```

```
df2.info()
```

```
df1.describe()
```

```
df2.describe()
```

```
df1.corr()
```

```
df2.corr()
```

```
sns.heatmap(df1.corr())
```

```
sns.heatmap(df2.corr())
```

```
df1['Year'].unique()
```

```
sns.pairplot(df1)
```

```
sns.pairplot(df2)
```

```
# Distribution of area dedicated to agriculture

sns.set_style('whitegrid')
sns.histplot(df1['AREA (1000 ha)'], bins=50, kde=True)
plt.title('Distribution of Area (1000 ha)')
plt.savefig('img1.png')
```

```
df1.head()
```

```
# Distribution of agricultural production

sns.set_style('whitegrid')
sns.histplot(df1['PRODUCTION (1000 tons)'], bins=50, kde=True)
plt.title('Distribution of Production (1000 tons)')
plt.savefig('img2.png')
```

```
df1.groupby('Year')['PRODUCTION (1000 tons)'].sum().plot(kind='line', marker='o')
plt.title('Total Agricultural Production Over the Years')
plt.xlabel('Year')
plt.ylabel('Total Production (1000 tons)')
plt.xticks(df1['Year'].unique().astype(int))
plt.xticks(rotation=90)
plt.grid(True)
plt.savefig('img3.png')
```

```
df1.groupby('Year')['YIELD (Tons per 1000 ha)'].mean().plot(kind='line', marker='o')
plt.title('Average Yield Over the Years')
plt.xlabel('Year')
plt.ylabel('Average Yield (Tons per 1000 ha)')
plt.xticks(df1['Year'].unique().astype(int))
plt.xticks(rotation=90)
plt.grid(True)
plt.savefig('img4.png')
```

```
sns.barplot(x='Crop Name', y='YIELD (Tons per 1000 ha)', data=df1)
plt.savefig('img5.png')
```

```
rice_data=df1[df1["Crop Name"]=="Rice"]
Wheat_data=df1[df1["Crop Name"]=="Wheat"]
Maize_data=df1[df1["Crop Name"]=="Maize"]
Chickpea_data=df1[df1["Crop Name"]=="Chickpea"]
Pigeonpea_data=df1[df1["Crop Name"]=="Pigeonpea"]
Sugarcane_data=df1[df1["Crop Name"]=="Sugarcane"]
Cotton_data=df1[df1["Crop Name"]=="Cotton"]
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=rice_data)
plt.title('Total Production of Rice Over the Years')
plt.xticks(rotation=45)
plt.savefig('img6.png')
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=Wheat_data)
plt.title('Total Production of Wheat Over the Years')
plt.xticks(rotation=45)
plt.savefig('img7.png')
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=Maize_data)
plt.title('Total Production of Maize Over the Years')
plt.xticks(rotation=45)
plt.savefig('img8.png')
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=Chickpea_data)
plt.title('Total Production of Chickpea Over the Years')
plt.xticks(rotation=45)
plt.savefig('img9.png')
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=Pigeonpea_data)
plt.title('Total Production of Pigeonpea Over the Years')
plt.xticks(rotation=45)
plt.savefig('img9.png')
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=Sugarcane_data)
plt.title('Total Production of Sugarcane Over the Years')
plt.xticks(rotation=45)
plt.savefig('img10.png')
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=Cotton_data)
plt.title('Total Production of Cotton Over the Years')
plt.xticks(rotation=45)
plt.savefig('img11.png')
```

```
dummy1 = pd.get_dummies(df1)
dummy1
```

```
from sklearn.model_selection import train_test_split

x = dummy1.drop(["PRODUCTION (1000 tons)","YIELD (Tons per 1000 ha)"], axis=1)
y = dummy1["PRODUCTION (1000 tons)"]

# Splitting data set - 25% test dataset and 75%

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.25, random_state=5)

print("x_train :",x_train.shape)
print("x_test :",x_test.shape)
print("y_train :",y_train.shape)
print("y_test :",y_test.shape)
```

```
print(x_train)
print(y_train)
```

```
# Training the Simple Linear Regression model .

from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(x_train,y_train)
```

```
# Predicting the test Results

lr_predict = model.predict(x_test)
lr_predict
```

```
model.score(x_test,y_test)
```

```
from sklearn.metrics import r2_score
r = r2_score(y_test,lr_predict)
print("R2 score : ",r)
```

```
plt.scatter(y_test,lr_predict)
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.title('Linear Regression')
```

```
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor(n_estimators = 11)
model.fit(x_train,y_train)
rf_predict = model.predict(x_test)
rf_predict
```

```
model.score(x_test,y_test)
```

```
# Calculating R2 score

from sklearn.metrics import r2_score
r1 = r2_score(y_test,rf_predict)
print("R2 score : ",r1)
```

```
# Calculating Adj. R2 score:

Adjr2_1 = 1 - (1-r)*(len(y_test)-1)/(len(y_test)-x_test.shape[1]-1)
print("Adj. R-Squared : {}".format(Adjr2_1))
```

```
ax = sns.distplot(y_test, hist = False, color = "r", label = "Actual value ")
sns.distplot(rf_predict, hist = False, color = "b", label = "Predicted Values", ax = ax)
plt.title('Random Forest Regression')
```

```
df2.head()
```

```
df2.rename(columns = {'label':'Crop'}, inplace = True)
```

```
df2['Crop'].unique()
```

```
sns.barplot(x="Crop", y="temperature", data=df2)
plt.title('Temperature for Different Crops')
plt.xticks(rotation = 90)
plt.savefig('img12.png')
```

```
sns.barplot(x="Crop", y="humidity", data=df2)
plt.title('Humidity for Different Crops')
plt.xticks(rotation = 90)
plt.savefig('img13.png')
```

```
sns.barplot(x="Crop", y="rainfall", data=df2)
plt.title('Rainfall for Different Crops')
plt.xticks(rotation = 90)
plt.savefig('img14.png')
```

```
sns.barplot(x="Crop", y="ph", data=df2)
plt.title('ph for Different Crops')
plt.xticks(rotation = 90)
plt.savefig('img15.png')
```

```
from sklearn.utils import shuffle

dummy2  = shuffle(df2,random_state=5)
dummy2.head()
```

```
# Selection of Feature and Target variables.

x = dummy2[['N', 'P','K','temperature', 'humidity', 'ph', 'rainfall']]
target = dummy2['Crop']
```

```
# Encoding target variable
y = pd.get_dummies(target)
y
```

```
# Splitting data set - 25% test dataset and 75%


from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.25, random_state= 0)

print("x_train :",x_train.shape)
print("x_test :",x_test.shape)
print("y_train :",y_train.shape)
print("y_test :",y_test.shape)
```

```
# Training

forest = RandomForestClassifier(random_state=1)
multi_target_forest = MultiOutputClassifier(forest, n_jobs=-1)
multi_target_forest.fit(x_train, y_train)
```

```
# Predicting test results

forest_pred = multi_target_forest.predict(x_test)
forest_pred
```

```
# Calculating Accuracy

from sklearn.metrics import accuracy_score
a1 = accuracy_score(y_test, forest_pred)
print('Accuracy score:', accuracy_score(y_test, forest_pred))
```

Crop Data Visualization

```
total_prod=df1.groupby('Dist Name')['PRODUCTION (1000 tons)'].sum()
total_prod=total_prod.sort_values(ascending=False)
total_prod.plot(kind='bar')
plt.xlabel('Dist Name')
plt.ylabel('Total Production (1000 tons)')
plt.xticks(rotation=90)
plt.grid(True)
```

```
total_yield=df1.groupby('Dist Name')['YIELD (Tons per 1000 ha)'].sum()
total_yield=total_yield.sort_values(ascending=False)
total_yield.plot(kind='bar')
plt.xlabel('Dist Name')
plt.ylabel('Total Yield (1000 tons)')
plt.xticks(rotation=90)
plt.grid(True)
```

```
prod_year=df1.groupby('Year')['PRODUCTION (1000 tons)'].sum()
prod_year=prod_year.sort_values(ascending=False)
prod_year.plot(kind='bar')
plt.title('Total Agricultural Production Over the Years')
plt.xlabel('Year')
plt.ylabel('Total Production (1000 tons)')

plt.xticks(rotation=90)
plt.grid(True)
```

```
df1.groupby('Year')['YIELD (Tons per 1000 ha)'].sum().plot(kind='bar')
plt.title('Total yield over the Years')
plt.xlabel('Year')
plt.ylabel('Yield')
```

```
max_area=df1.groupby('Dist Name')['AREA (1000 ha)'].sum()
max_area=max_area.sort_values(ascending=False)
max_area.plot(kind='bar')
plt.title('Area dedicated to agriculture in districts')
plt.xlabel('District')
plt.ylabel('Area (1000 ha)')
plt.grid(True)
```

Rice

```
df1[df1['Crop Name']=='Rice'].groupby('Dist Name')['PRODUCTION (1000 tons)'].sum().plot(kind='bar')
plt.title('Total Production of rice in districts')
plt.xlabel('Dist Name')
plt.ylabel('Production')
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=df1[df1['Crop Name']=='Rice'])
plt.title("Production of Rice over the years")
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='YIELD (Tons per 1000 ha)', data=df1[df1['Crop Name']=='Rice'])
plt.title("Yield of Rice over the years")
plt.xticks(rotation=90)
```

Wheat

```
df1[df1['Crop Name']=='Wheat'].groupby('Dist Name')['PRODUCTION (1000 tons)'].sum().plot(kind='bar')
plt.title('Total Production of wheat in districts')
plt.xlabel('Dist Name')
plt.ylabel('Production')
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=df1[df1['Crop Name']=='Wheat'])
plt.title("Production of Wheat over the years")
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='YIELD (Tons per 1000 ha)', data=df1[df1['Crop Name']=='Wheat'])
plt.title("Yield of Wheat over the years")
plt.xticks(rotation=90)
```

Maize

```
df1[df1['Crop Name']=='Maize'].groupby('Dist Name')['PRODUCTION (1000 tons)'].sum().plot(kind='bar')
plt.title('Total Production of Maize in districts')
plt.xlabel('Dist Name')
plt.ylabel('Production')
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=df1[df1['Crop Name']=='Maize'])
plt.title("Production of Maize over the years")
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='YIELD (Tons per 1000 ha)', data=df1[df1['Crop Name']=='Maize'])
plt.title("Yield of Maize over the years")
plt.xticks(rotation=90)
```

Chickpea

```
df1[df1['Crop Name']=='Chickpea'].groupby('Dist Name')['PRODUCTION (1000 tons)'].sum().plot(kind='bar')
plt.title('Total Production of Chickpea in districts')
plt.xlabel('Dist Name')
plt.ylabel('Production')
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=df1[df1['Crop Name']=='Chickpea'])
plt.title("Production of Chickpea over the years")
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='YIELD (Tons per 1000 ha)', data=df1[df1['Crop Name']=='Chickpea'])
plt.title("Yield of Chickpea over the years")
plt.xticks(rotation=90)
```

Pigeonpea

```
df1[df1['Crop Name']=='Pigeonpea'].groupby('Dist Name')['PRODUCTION (1000 tons)'].sum().plot(kind='bar')
plt.title('Total Production of Pigeonpea in districts')
plt.xlabel('Dist Name')
plt.ylabel('Production')
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=df1[df1['Crop Name']=='Pigeonpea'])
plt.title("Production of Pigeonpea over the years")
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='YIELD (Tons per 1000 ha)', data=df1[df1['Crop Name']=='Pigeonpea'])
plt.title("Yield of Pigeonpea over the years")
plt.xticks(rotation=90)
```

Sugarcane

```
df1[df1['Crop Name']=='Sugarcane'].groupby('Dist Name')['PRODUCTION (1000 tons)'].sum().plot(kind='bar')
plt.title('Total Production of sugarcane in districts')
plt.xlabel('Dist Name')
plt.ylabel('Production')
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=df1[df1['Crop Name']=='Sugarcane'])
plt.title("Production of Sugarcane over the years")
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='YIELD (Tons per 1000 ha)', data=df1[df1['Crop Name']=='Sugarcane'])
plt.title("Yield of Sugarcane over the years")
plt.xticks(rotation=90)
```

Cotton

```
df1[df1['Crop Name']=='Cotton'].groupby('Dist Name')['PRODUCTION (1000 tons)'].sum().plot(kind='bar')
plt.title('Total Production of Cotton in districts')
plt.xlabel('Dist Name')
plt.ylabel('Production')
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='PRODUCTION (1000 tons)', data=df1[df1['Crop Name']=='Cotton'])
plt.title("Production of Cotton over the years")
plt.xticks(rotation=90)
```

```
sns.barplot(x='Year', y='YIELD (Tons per 1000 ha)', data=df1[df1['Crop Name']=='Cotton'])
plt.title("Yield of Cotton over the years")
plt.xticks(rotation=90)
```

```
sns.set(style='whitegrid')
plt.figure(figsize=(20, 20))

plt.subplot(3, 3, 1)
sns.histplot(df2['N'], kde=True, bins=30, color='blue')

plt.subplot(3, 3, 2)
sns.histplot(df2['P'], kde=True, bins=30, color='Orange')

plt.subplot(3, 3, 3)
sns.histplot(df2['K'], kde=True, bins=30, color='green')

plt.subplot(3, 3, 4)
sns.histplot(df2['temperature'], kde=True, bins=30, color='red')

plt.subplot(3, 3, 5)
sns.histplot(df2['humidity'], kde=True, bins=30, color='brown')

plt.subplot(3, 3, 6)
sns.histplot(df2['ph'], kde=True, bins=30, color='yellow')

plt.subplot(3, 3, 7)
sns.histplot(df2['rainfall'], kde=True, bins=30, color='purple')
```

APPENDIX B

SCREENSHOTS

Crop Recommendation Outputs

Crop Yield and Production Dataset (df1)

	Dist Code	Year	State Code	State Name	Dist Name	Crop Name	AREA (1000 ha)	PRODUCTION (1000 tons)	YIELD (Tons per 1000 ha)
0	44	2000	1	Andhra Pradesh	Srikakulam	Rice	274.60	511.34	1.862127
1	44	2001	1	Andhra Pradesh	Srikakulam	Rice	235.44	502.11	2.132645
2	44	2002	1	Andhra Pradesh	Srikakulam	Rice	201.50	328.50	1.630273
3	44	2003	1	Andhra Pradesh	Srikakulam	Rice	246.96	545.82	2.210155
4	44	2004	1	Andhra Pradesh	Srikakulam	Rice	256.78	601.96	2.344264
...	...	...	...	...	...	...	...	...	...
1381	54	2013	1	Andhra Pradesh	Chittoor	Cotton	0.35	0.16	0.457143
1382	54	2014	1	Andhra Pradesh	Chittoor	Cotton	0.85	0.36	0.423529
1383	54	2015	1	Andhra Pradesh	Chittoor	Cotton	1.14	0.36	0.315789
1384	54	2016	1	Andhra Pradesh	Chittoor	Cotton	0.78	0.31	0.397436
1385	54	2017	1	Andhra Pradesh	Chittoor	Cotton	0.50	0.24	0.480000

1386 rows × 9 columns

Required Parameters Dataset (df2)

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
...	...	...	...	...	...	...	...	...
2195	107	34	32	26.774637	66.413269	6.780064	177.774507	coffee
2196	99	15	27	27.417112	56.636362	6.086922	127.924610	coffee
2197	118	33	30	24.131797	67.225123	6.362608	173.322839	coffee
2198	117	32	34	26.272418	52.127394	6.758793	127.175293	coffee
2199	104	18	30	23.603016	60.396475	6.779833	140.937041	coffee

2200 rows × 8 columns

Column Name df1

```
Index(['Dist Code', 'Year', 'State Code', 'State Name', 'Dist Name',
       'Crop Name', 'AREA (1000 ha)', 'PRODUCTION (1000 tons)',
       'YIELD (Tons per 1000 ha)'],
      dtype='object')
```

Column Name df2

```
Index(['N', 'P', 'K', 'temperature', 'humidity', 'ph', 'rainfall', 'label'], dtype='object')
```

Crops Selected df1

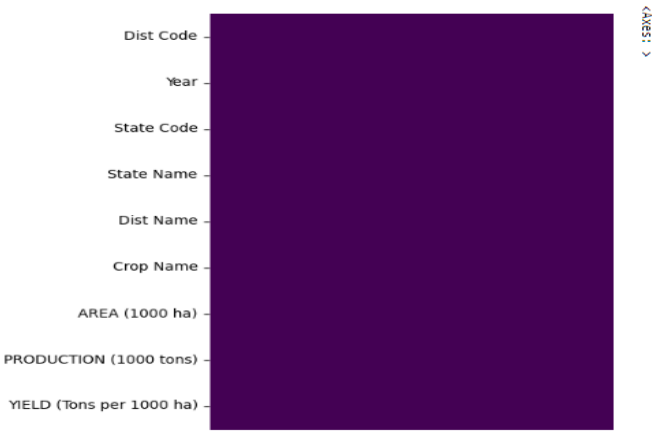
```
array(['Rice', 'Wheat', 'Maize', 'Chickpea', 'Pigeonpea', 'Sugarcane',
       'Cotton'], dtype=object)
```

```
Rice         198
Wheat        198
Maize        198
Chickpea     198
Pigeonpea    198
Sugarcane    198
Cotton       198
Name: Crop Name, dtype: int64
```

Data Cleaning df1

	Dist Code	Year	State Code	State Name	Dist Name	Crop Name	AREA (1000 ha)	PRODUCTION (1000 tons)	YIELD (Tons per 1000 ha)
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
1381	False	False	False	False	False	False	False	False	False
1382	False	False	False	False	False	False	False	False	False
1383	False	False	False	False	False	False	False	False	False
1384	False	False	False	False	False	False	False	False	False
1385	False	False	False	False	False	False	False	False	False

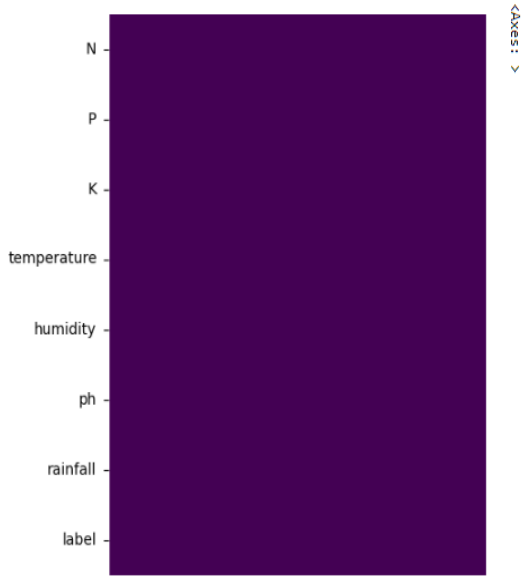
1386 rows × 9 columns



Data cleaning df2

	N	P	K	temperature	humidity	ph	rainfall	label
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
2195	False	False	False	False	False	False	False	False
2196	False	False	False	False	False	False	False	False
2197	False	False	False	False	False	False	False	False
2198	False	False	False	False	False	False	False	False
2199	False	False	False	False	False	False	False	False

2200 rows × 8 columns



The df.**info()** method provides a concise summary of a DataFrame, displaying information such as the data types, non-null counts, and memory usage.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1386 entries, 0 to 1385
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Dist Code                 1386 non-null   int64
 1   Year                      1386 non-null   int64
 2   State Code                1386 non-null   int64
 3   State Name                1386 non-null   object
 4   Dist Name                 1386 non-null   object
 5   Crop Name                 1386 non-null   object
 6   AREA (1000 ha)            1386 non-null   float64
 7   PRODUCTION (1000 tons)    1386 non-null   float64
 8   YIELD (Tons per 1000 ha)  1386 non-null   float64
dtypes: float64(3), int64(3), object(3)
memory usage: 97.6+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2200 entries, 0 to 2199
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   N            2200 non-null   int64
 1   P            2200 non-null   int64
 2   K            2200 non-null   int64
 3   temperature  2200 non-null   float64
 4   humidity     2200 non-null   float64
 5   ph           2200 non-null   float64
 6   rainfall     2200 non-null   float64
 7   label        2200 non-null   object
dtypes: float64(4), int64(3), object(1)
memory usage: 137.6+ KB
```

The **df.describe()** method in pandas is used to generate descriptive statistics of a DataFrame, including measures of central tendency, dispersion, and shape of the distribution of a dataset

	Dist Code	Year	State Code	AREA (1000 ha)	PRODUCTION (1000 tons)	YIELD (Tons per 1000 ha)
count	1386.000000	1386.00	1386.0	1386.000000	1386.000000	1386.000000
mean	49.000000	2008.50	1.0	50.249784	142.235678	2.475431
std	3.163419	5.19	0.0	92.368885	311.600282	2.718146
min	44.000000	2000.00	1.0	0.000000	0.000000	0.000000
25%	46.000000	2004.00	1.0	0.500000	0.610000	0.339237
50%	49.000000	2008.50	1.0	11.280000	11.765000	1.372017
75%	52.000000	2013.00	1.0	48.405000	118.442500	3.924861
max	54.000000	2017.00	1.0	471.670000	1693.910000	13.115385

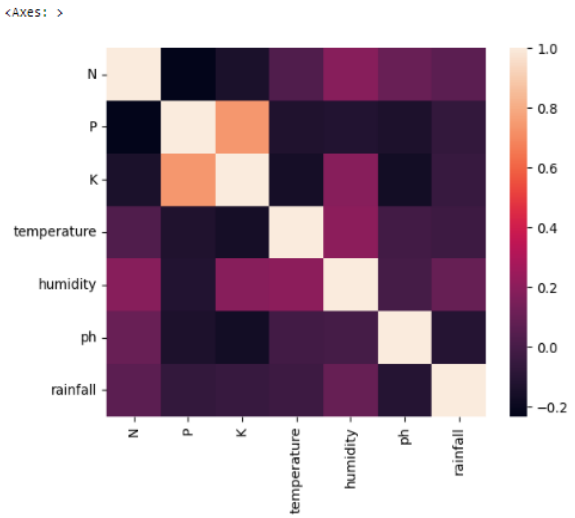
	N	P	K	temperature	humidity	ph	rainfall
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	50.551818	53.362727	48.149091	25.616244	71.481779	6.469480	103.463655
std	36.917334	32.985883	50.647931	5.063749	22.263812	0.773938	54.958389
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	20.211267
25%	21.000000	28.000000	20.000000	22.769375	60.261953	5.971693	64.551686
50%	37.000000	51.000000	32.000000	25.598693	80.473146	6.425045	94.867624
75%	84.250000	68.000000	49.000000	28.561654	89.948771	6.923643	124.267508
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	298.560117

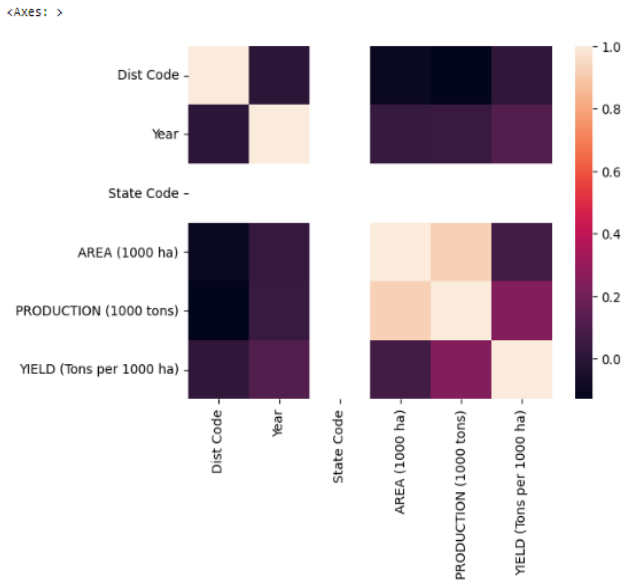
The **df.corr()** method in pandas is used to compute pairwise correlation of columns, excluding NA/null values. It calculates the correlation coefficients between all pairs of numeric columns in a DataFrame. The correlation coefficient ranges from -1 to 1, where:

- 1 indicates a perfect positive correlation,
- -1 indicates a perfect negative correlation,
- 0 indicates no correlation.

	Dist Code	Year	State Code	AREA (1000 ha)	PRODUCTION (1000 tons)	YIELD (Tons per 1000 ha)
Dist Code	1.000000e+00	-1.564597e-14	NaN	-0.093788	-0.128830	0.012428
Year	-1.564597e-14	1.000000e+00	NaN	0.030703	0.040291	0.110288
State Code	NaN	NaN	NaN	NaN	NaN	NaN
AREA (1000 ha)	-9.378828e-02	3.070346e-02	NaN	1.000000	0.917218	0.071117
PRODUCTION (1000 tons)	-1.288302e-01	4.029108e-02	NaN	0.917218	1.000000	0.242992
YIELD (Tons per 1000 ha)	1.242821e-02	1.102876e-01	NaN	0.071117	0.242992	1.000000

	N	P	K	temperature	humidity	ph	rainfall
N	1.000000	-0.231460	-0.140512	0.026504	0.190688	0.096683	0.059020
P	-0.231460	1.000000	0.736232	-0.127541	-0.118734	-0.138019	-0.063839
K	-0.140512	0.736232	1.000000	-0.160387	0.190859	-0.169503	-0.053461
temperature	0.026504	-0.127541	-0.160387	1.000000	0.205320	-0.017795	-0.030084
humidity	0.190688	-0.118734	0.190859	0.205320	1.000000	-0.008483	0.094423
ph	0.096683	-0.138019	-0.169503	-0.017795	-0.008483	1.000000	-0.109069
rainfall	0.059020	-0.063839	-0.053461	-0.030084	0.094423	-0.109069	1.000000



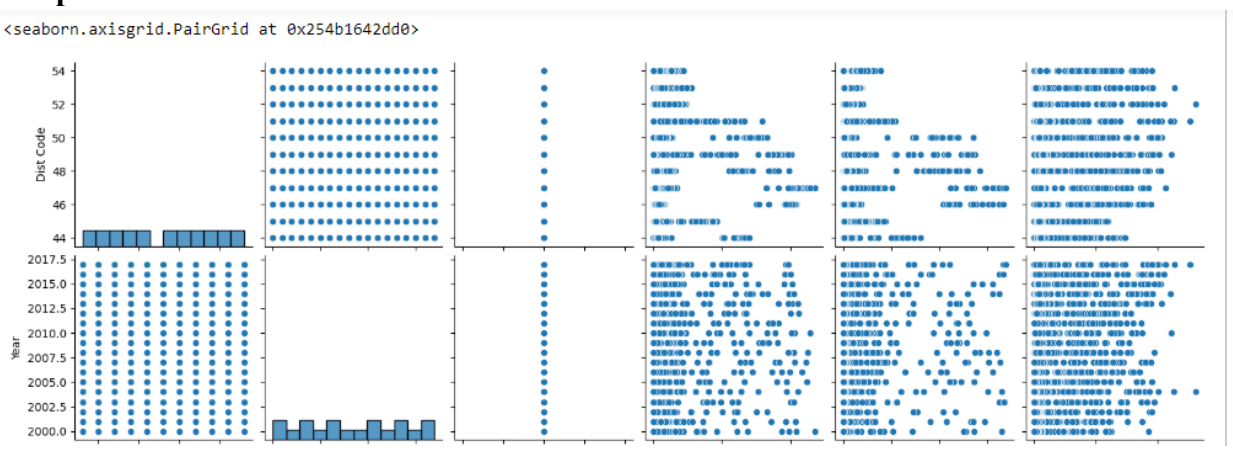


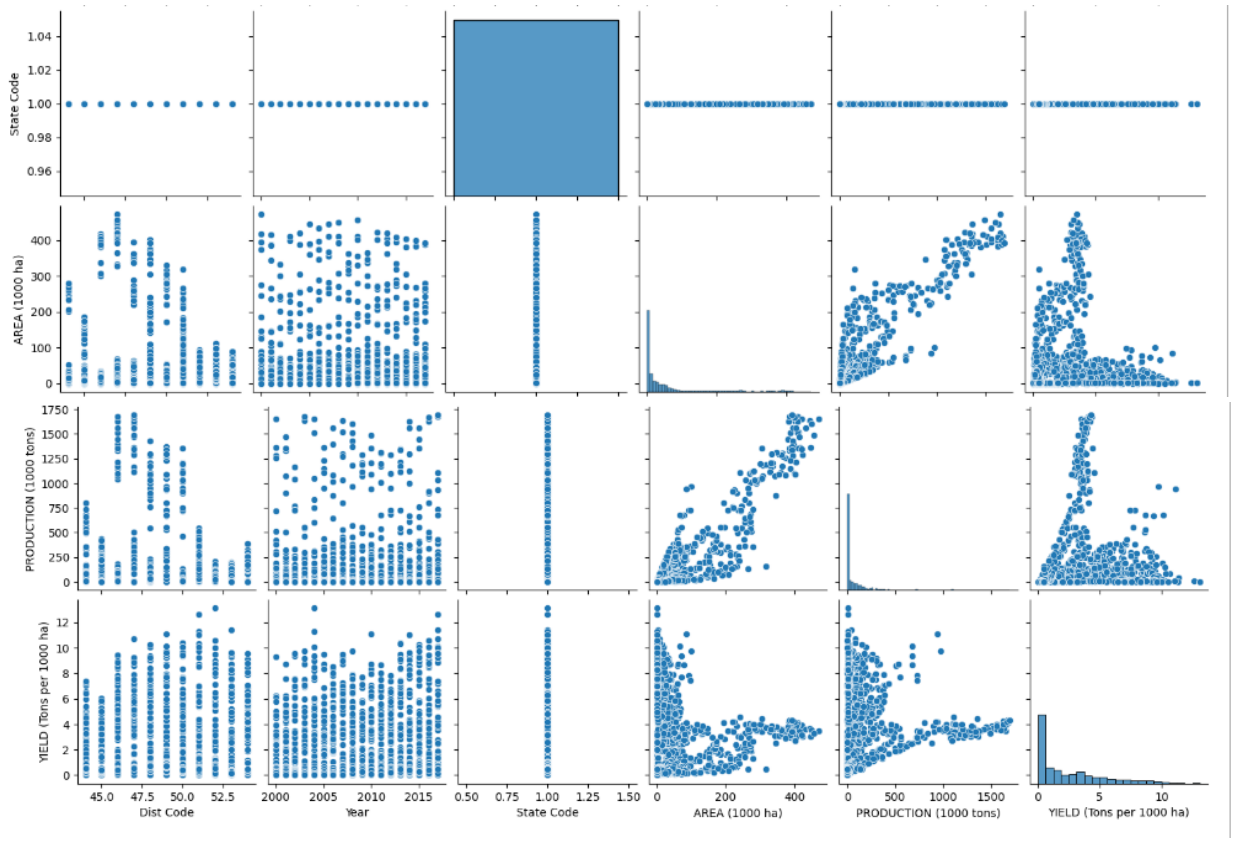
Year Chosen

```
array([2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010,
       2011, 2012, 2013, 2014, 2015, 2016, 2017], dtype=int64)
```

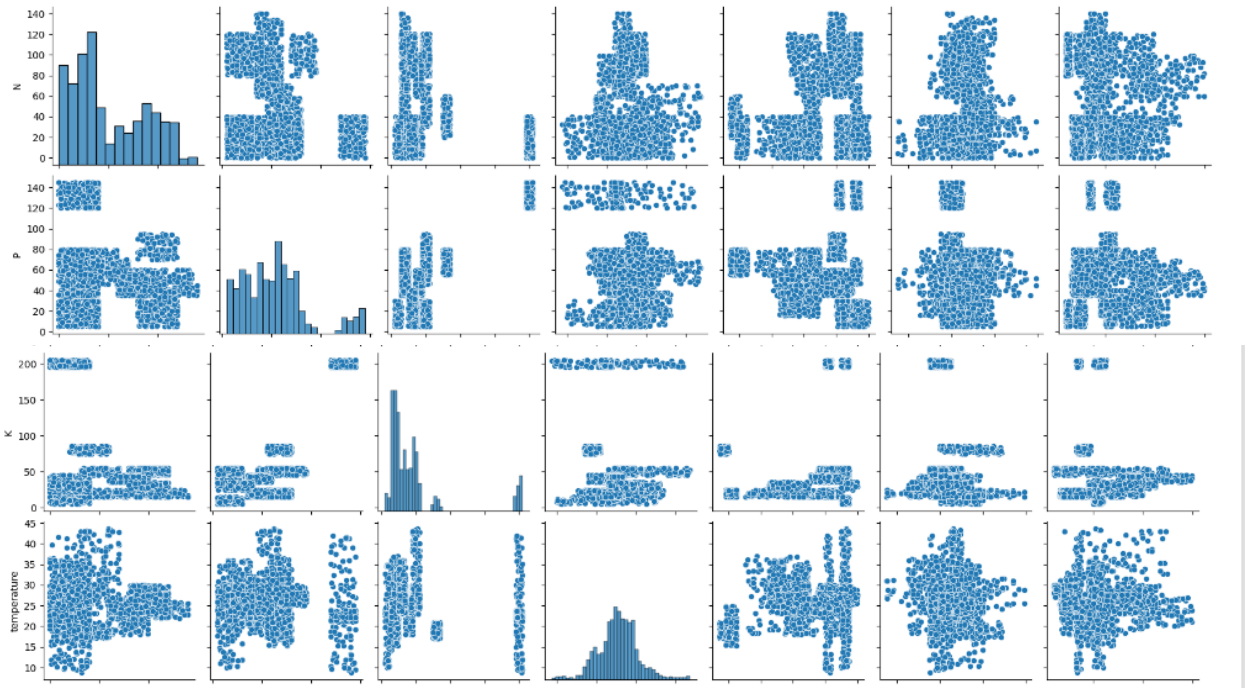
Pairplot

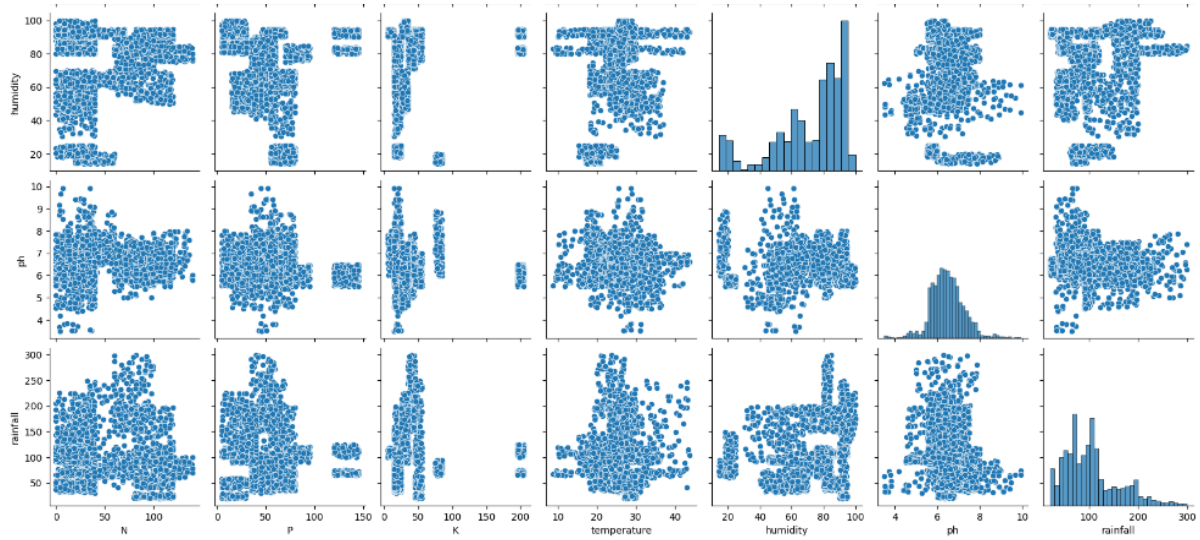
```
<seaborn.axisgrid.PairGrid at 0x254b1642dd0>
```



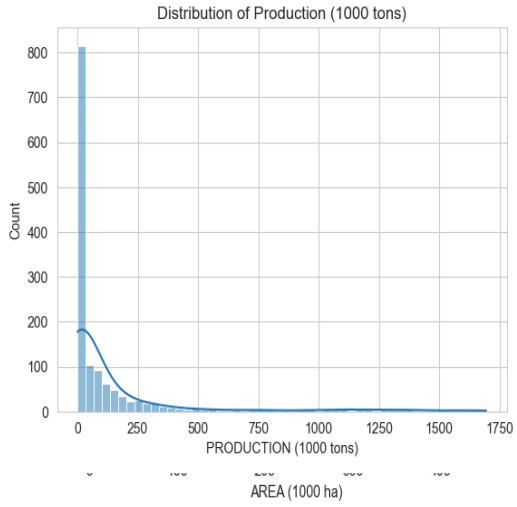


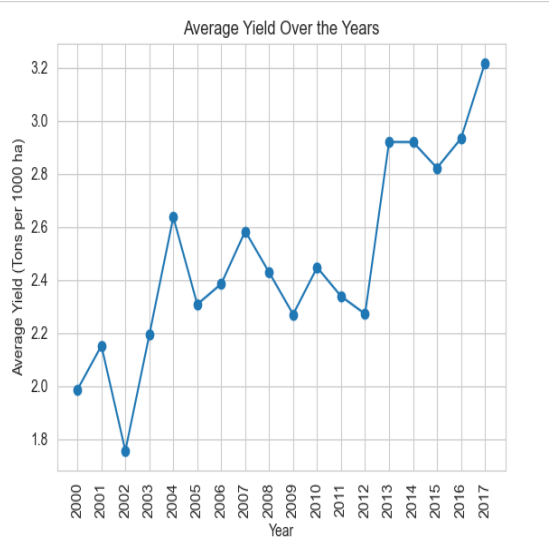

```
<seaborn.axisgrid.PairGrid at 0x254c74b4040>
```

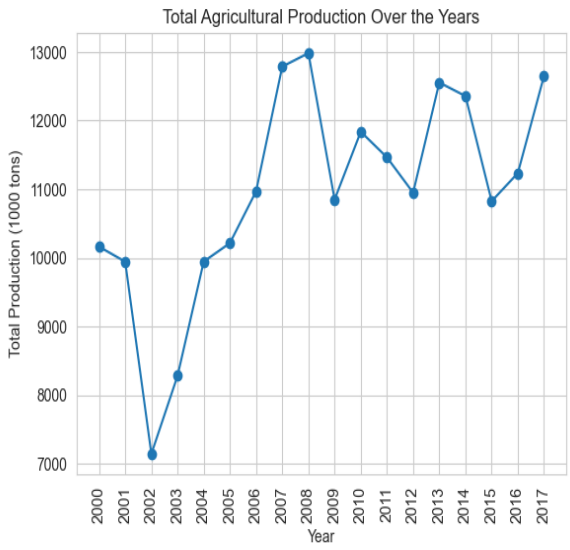


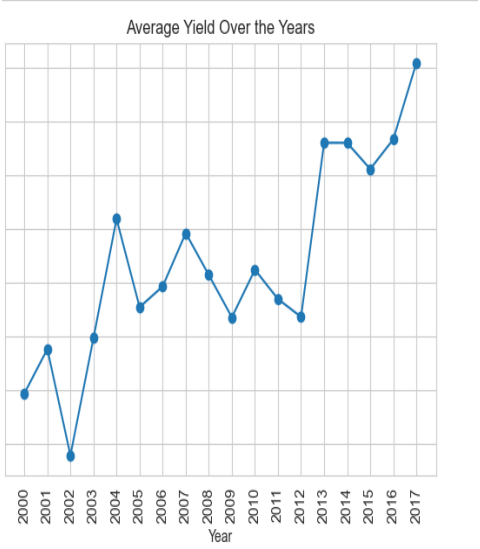


	Dist Code	Year	State Code	State Name	Dist Name	Crop Name	AREA (1000 ha)	PRODUCTION (1000 tons)	YIELD (Tons per 1000 ha)
0	44	2000	1	Andhra Pradesh	Srikakulam	Rice	274.60	511.34	1.862127
1	44	2001	1	Andhra Pradesh	Srikakulam	Rice	235.44	502.11	2.132645
2	44	2002	1	Andhra Pradesh	Srikakulam	Rice	201.50	328.50	1.630273
3	44	2003	1	Andhra Pradesh	Srikakulam	Rice	246.96	545.82	2.210155
4	44	2004	1	Andhra Pradesh	Srikakulam	Rice	256.78	601.96	2.344264

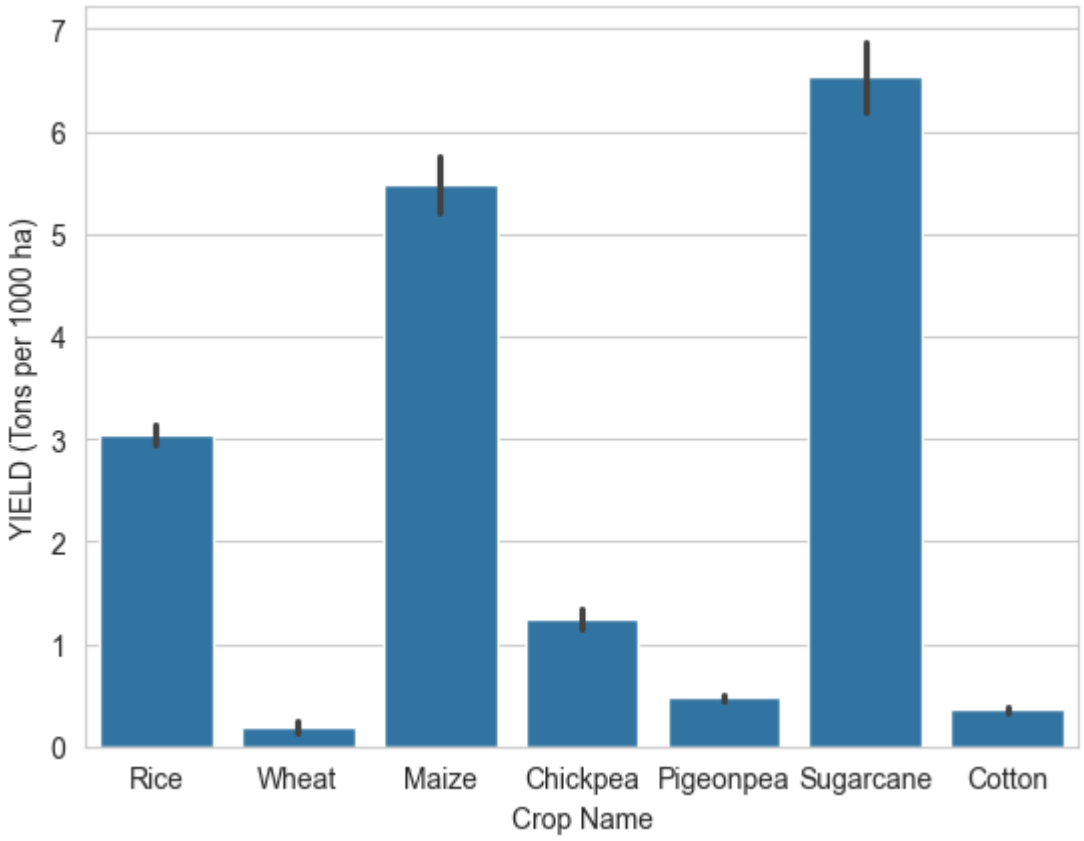




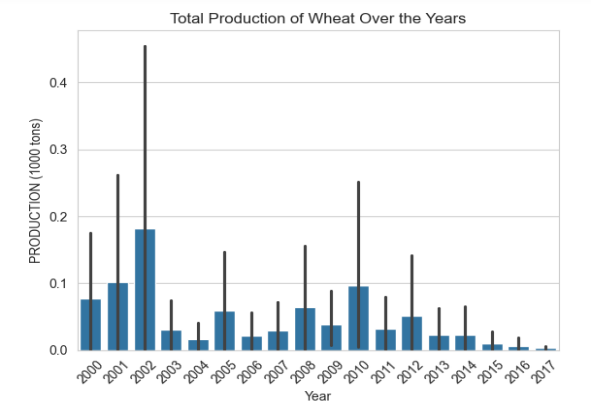

Total Agricultural Production Over the Years
Total Production (1000 tons)
13000
12000
11000
10000
9000
8000
7000
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year

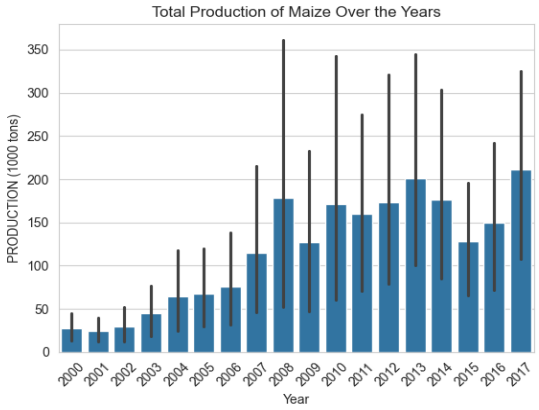

Average Yield Over the Years
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
Year

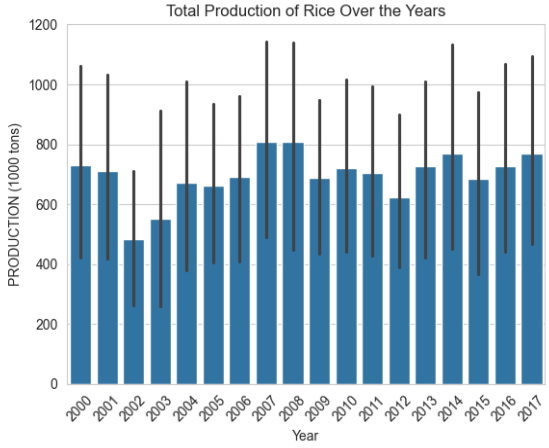
Crop Yields

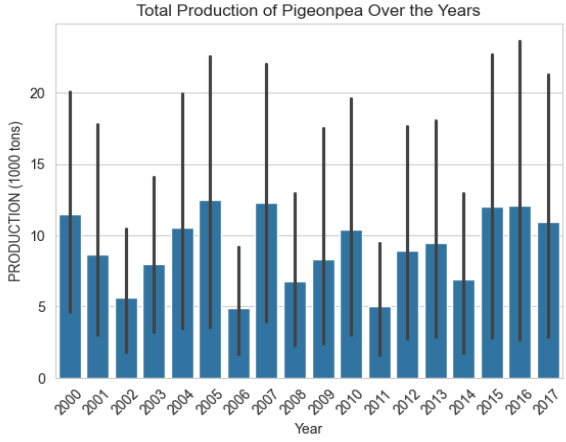

YIELD (Tons per 1000 ha)
7
6
5
4
3
2
1
0
Rice
Wheat
Maize
Chickpea
Pigeonpea
Sugarcane
Cotton
Crop Name

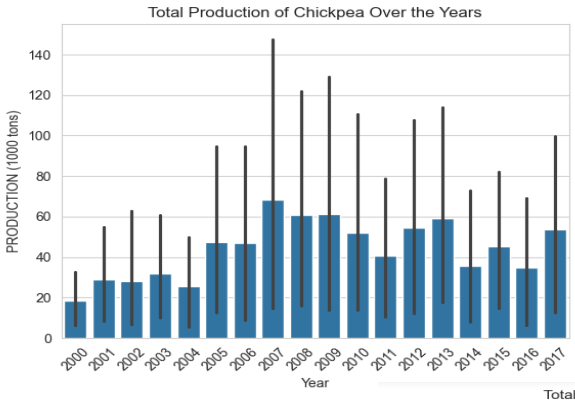
Crop Graphs

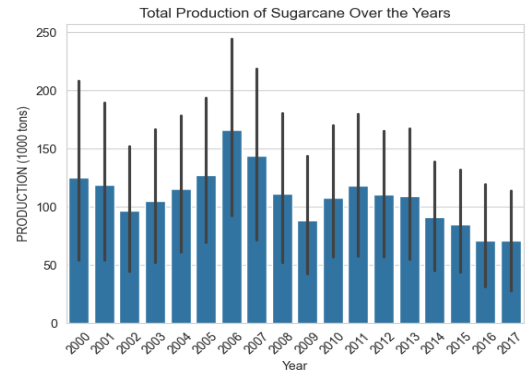

Total Production of Wheat Over the Years
PRODUCTION (1000 tons)
Year

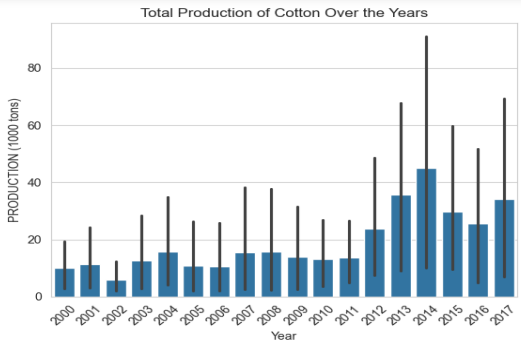

Total Production of Maize Over the Years
PRODUCTION (1000 tons)
Year


Total Production of Rice Over the Years
PRODUCTION (1000 tons)
Year


Total Production of Pigeonpea Over the Years
PRODUCTION (1000 tons)
Year


Total Production of Chickpea Over the Years
PRODUCTION (1000 tons)
Year


Total Production of Sugarcane Over the Years
PRODUCTION (1000 tons)
Year


Total Production of Cotton Over the Years
PRODUCTION (1000 tons)
Year

The get_dummies function has been applied to the dataset, converting categorical variables into dummy/indicator variables. This is typically done to prepare the data for machine learning models that require numerical input. The head of the transformed dataframe shows the new binary columns for state names, district names, and crop names, alongside the original numerical columns.

	Dist Code	Year	State Code	AREA (1000 ha)	PRODUCTION (1000 tons)	YIELD (Tons per 1000 ha)	State Name_Andhra Pradesh	Dist Name_Ananthapur	Dist Name_Chittoor	Dist Name_East Godavari	...	Dist Name_Srikakulam	Name_Visakhapa
0	44	2000	1	274.60	511.34	1.862127	1	0	0	0	...	1	
1	44	2001	1	235.44	502.11	2.132645	1	0	0	0	...	1	
2	44	2002	1	201.50	328.50	1.630273	1	0	0	0	...	1	
3	44	2003	1	246.96	545.82	2.210155	1	0	0	0	...	1	
4	44	2004	1	256.78	601.96	2.344264	1	0	0	0	...	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
1381	54	2013	1	0.35	0.16	0.457143	1	0	1	0	...	0	
1382	54	2014	1	0.85	0.36	0.423529	1	0	1	0	...	0	
1383	54	2015	1	1.14	0.36	0.315789	1	0	1	0	...	0	
1384	54	2016	1	0.78	0.31	0.397436	1	0	1	0	...	0	
1385	54	2017	1	0.50	0.24	0.480000	1	0	1	0	...	0	

1386 rows × 25 columns

Training and Testing

```
x_train : (1039, 23)
x_test : (347, 23)
y_train : (1039,)
y_test : (347,)
```

```
      Dist Code  Year  State Code  AREA (1000 ha)  State Name_Andhra Pradesh  \
667          48  2001           1            0.03                          1
615          45  2003           1            0.31                          1
358          52  2016           1            0.00                          1
830          46  2002           1            1.00                          1
1248         47  2006           1            4.37                          1
...         ...   ...         ...             ...                        ...
73           48  2001           1          343.59                          1
1142         52  2008           1            0.19                          1
998          44  2008           1           21.40                          1
206          44  2008           1            0.00                          1
867          48  2003           1            9.63                          1

      Dist Name_Ananthapur  Dist Name_Chittoor  Dist Name_East Godavari  \
667                      0                   0                        0
615                      0                   0                        0
358                      1                   0                        0
830                      0                   0                        1
1248                     0                   0                        0
...                    ...                 ...                      ...
73                       0                   0                        0
1142                     1                   0                        0
998                      0                   0                        0
206                      0                   0                        0
867                      0                   0                        0
```

```
      Dist Name_Guntur  Dist Name_Kadapa YSR  ...  Dist Name_Srikakulam  \
667                  0                     0  ...                     0
615                  0                     0  ...                     0
358                  0                     0  ...                     0
830                  0                     0  ...                     0
1248                 0                     0  ...                     0
...                ...                   ...  ...                   ...
73                   0                     0  ...                     0
1142                 0                     0  ...                     0
998                  0                     0  ...                     1
206                  0                     0  ...                     1
867                  0                     0  ...                     0

      Dist Name_Visakhapatnam  Dist Name_West Godavari  Crop Name_Chickpea  \
667                         0                        0                   1
615                         1                        0                   1
358                         0                        0                   0
830                         0                        0                   0
1248                        0                        1                   0
...                       ...                      ...                 ...
73                          0                        0                   0
1142                        0                        0                   0
998                         0                        0                   0
206                         0                        0                   0
867                         0                        0                   0

     Crop Name_Cotton  Crop Name_Maize  Crop Name_Pigeonpea  Crop Name_Rice  \
667                 0                0                    0               0
615                 0                0                    0               0
358                 0                0                    0               0
830                 0                0                    1               0
1248                1                0                    0               0
...               ...              ...                  ...             ...
73                  0                0                    0               1
1142                0                0                    0               0
998                 0                0                    0               0
206                 0                0                    0               0
867                 0                0                    1               0

       Crop Name_Sugarcane  Crop Name_Wheat
667                      0                0
615                      0                0
358                      0                1
830                      0                0
1248                     0                0
...                    ...              ...
73                       0                0
1142                     1                0
998                      1                0
206                      0                1
867                      0                0

[1039 rows x 23 columns]
667        0.06
615        0.55
358        0.00
830        0.00
1248       2.29
          ...
73      1073.14
1142       1.23
998       98.86
206        0.00
867        6.36
Name: PRODUCTION (1000 tons), Length: 1039, dtype: float64
```

Linear Regression

```
▾ LinearRegression
LinearRegression()
```

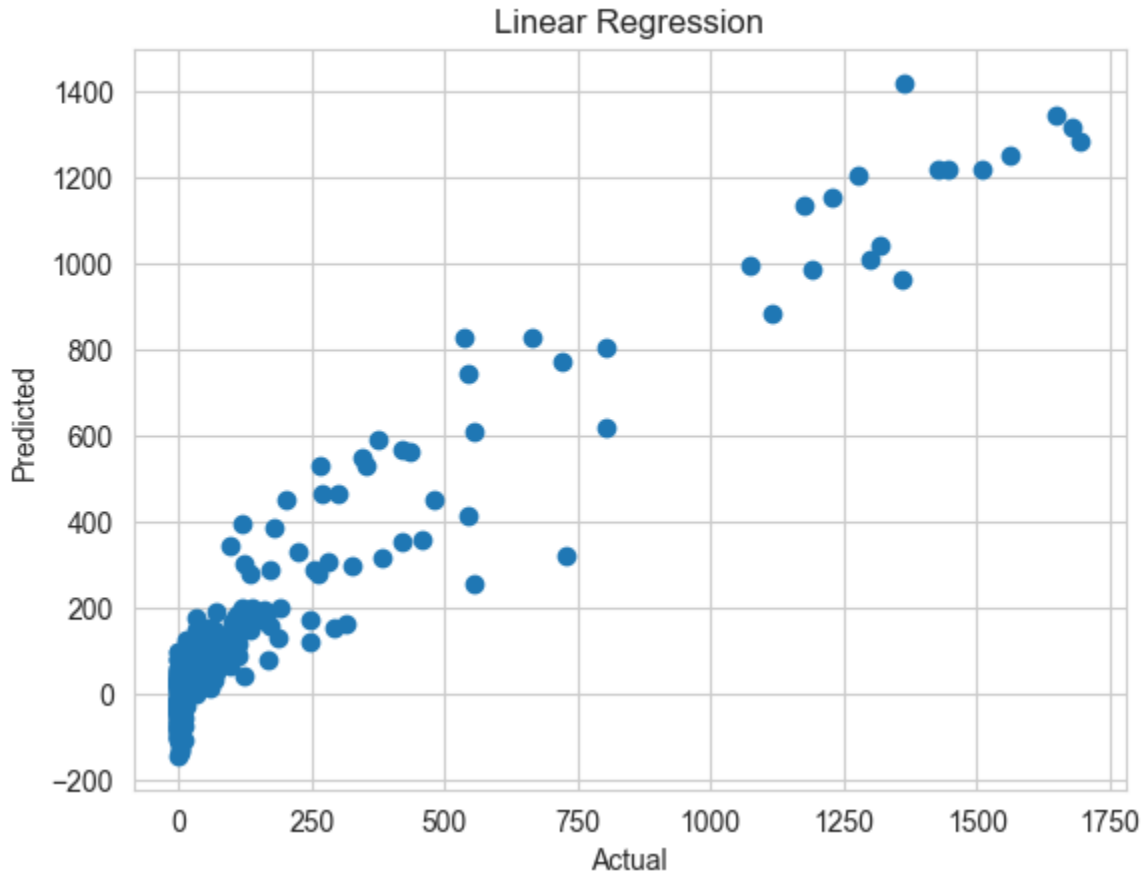
```
0.907255314613902
```

The model's score on the test set is approximately 0.907, which indicates a high level of accuracy in the predictions relative to the true values. This score is the coefficient of determination, also known as R-squared, which measures the proportion of variance in the dependent variable that is predictable from the independent variables.

```
R2 score :  0.907255314613902
```

The R-squared score for the model's predictions on the test set is approximately 0.907, indicating a strong correlation between the predicted values and the actual values.

```
Text(0.5, 1.0, 'Linear Regression')
```



The scatter plot shows a strong linear relationship between the actual and predicted values, indicating that the dataset is suitable for linear regression. The points are closely clustered around the diagonal line, suggesting that the model's predictions align well with the true values. Therefore, the dataset is well-suited for linear regression analysis.

Random Forest Regression

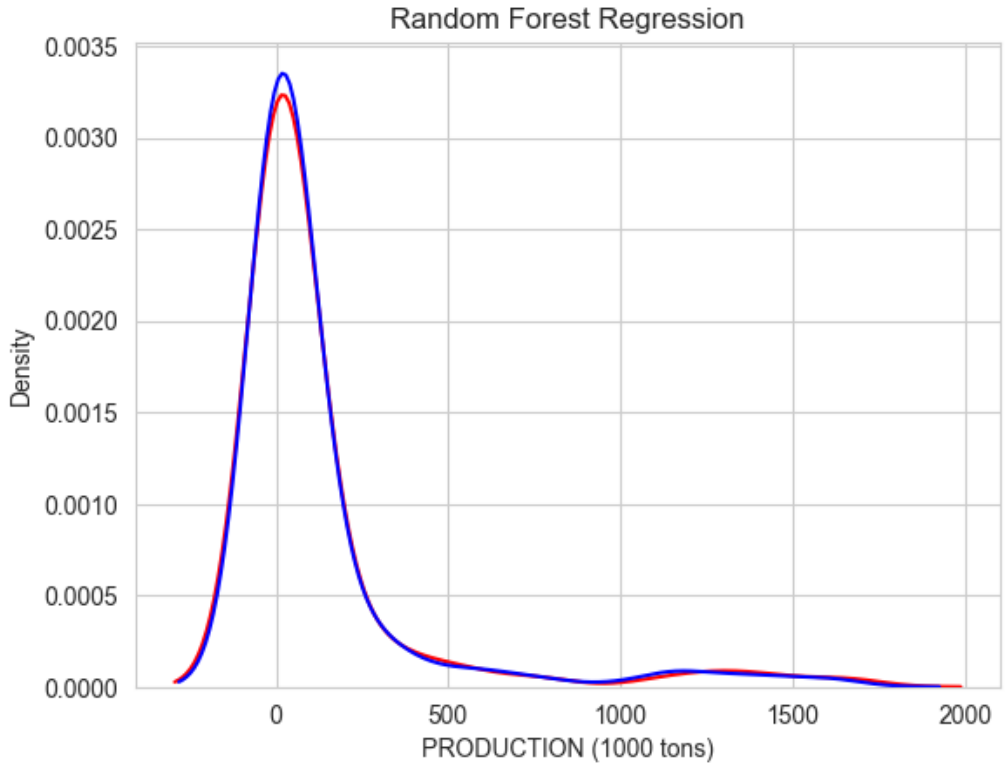
```
0.9801146542225873
```

The score of the RandomForestRegressor model on the test set is approximately 0.985, indicating a very high level of accuracy in the predictions relative to the true values. This score is significantly higher than the score obtained from the linear regression model, suggesting that the random forest model is better suited for this dataset.

```
R2 score :  0.9801146542225873
```

```
Adj. R-Squared : 0.9006512038898145
```

```
Text(0.5, 1.0, 'Random Forest Regression')
```



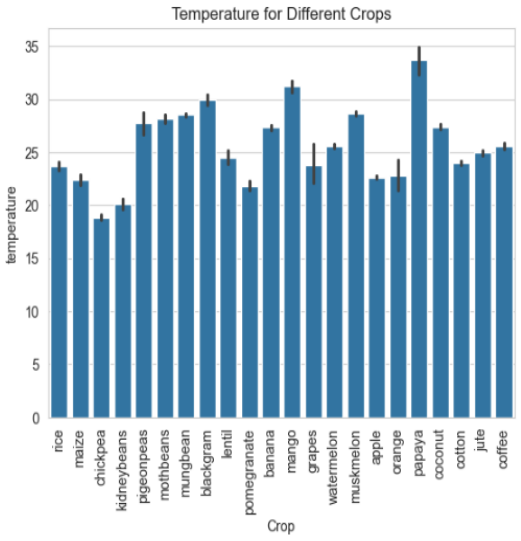
The distribution plot shows the density of the actual and predicted values from the Random Forest Regression model. The overlap between the two distributions indicates the model's accuracy in predicting the test data

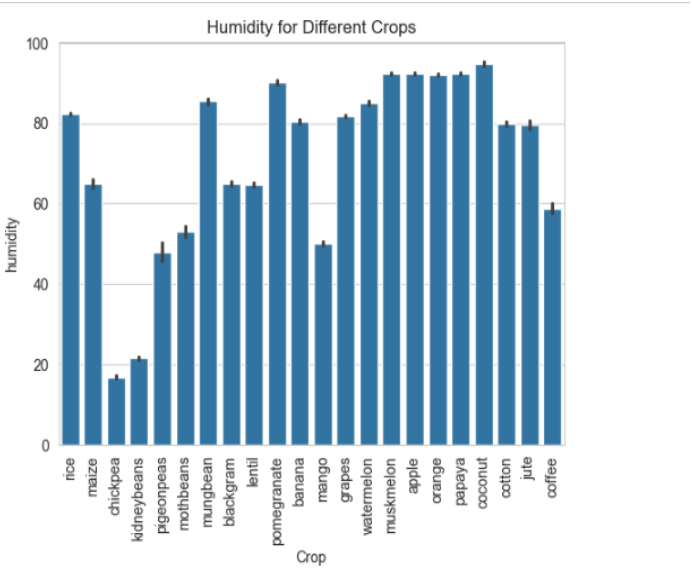
Prediction model for df2

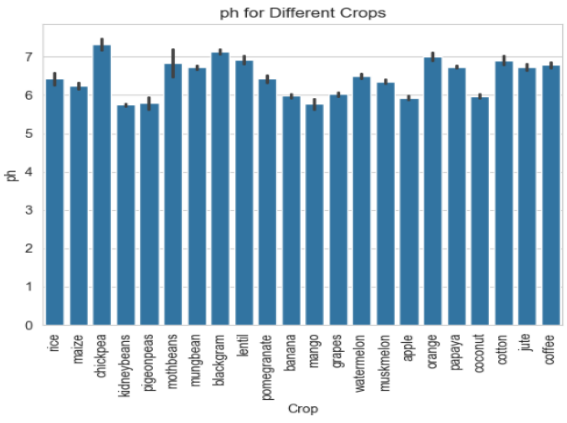
	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

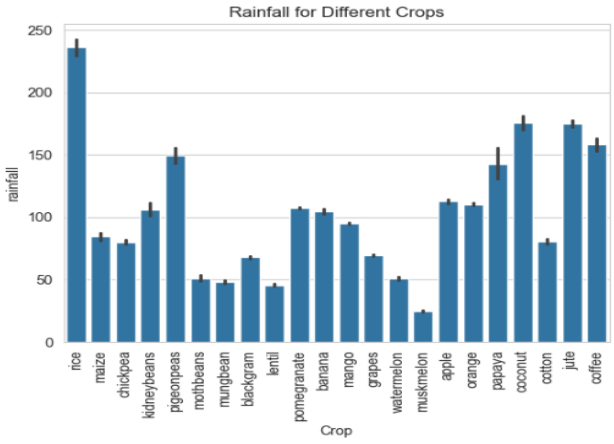
Crops

```
array(['rice', 'maize', 'chickpea', 'kidneybeans', 'pigeonpeas',
       'mothbeans', 'mungbean', 'blackgram', 'lentil', 'pomegranate',
       'banana', 'mango', 'grapes', 'watermelon', 'muskmelon', 'apple',
       'orange', 'papaya', 'coconut', 'cotton', 'jute', 'coffee'],
      dtype=object)
```









Shuffled Data

	N	P	K	temperature	humidity	ph	rainfall	Crop
1270	6	140	205	17.665584	82.929034	6.313086	69.867126	grapes
1481	98	22	47	29.072653	91.915332	6.341401	28.835684	muskmelon
1832	38	14	30	26.924495	91.201060	5.570745	194.902214	coconut
293	35	63	76	17.815645	17.607566	7.714153	90.820976	chickpea
1307	85	22	53	25.965342	89.770767	6.849472	59.463386	watermelon

Dummy Data

	apple	banana	blackgram	chickpea	coconut	coffee	cotton	grapes	jute	kidneybeans	...	mango	mothbeans	mungbean	muskmelon	orange	papay
1270	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	
1481	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	
1832	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	
293	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	
1307	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
740	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	
1032	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2121	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	
1424	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	
1725	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

2200 rows × 22 columns

MultiOutput Classifier

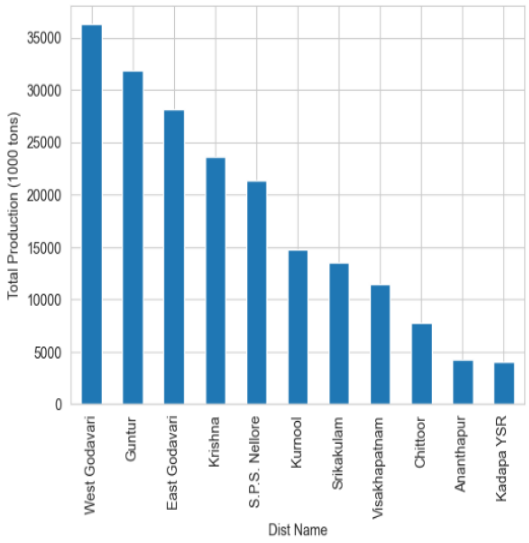
```
MultiOutputClassifier(estimator=RandomForestClassifier(random_state=1),
                      n_jobs=-1)
```

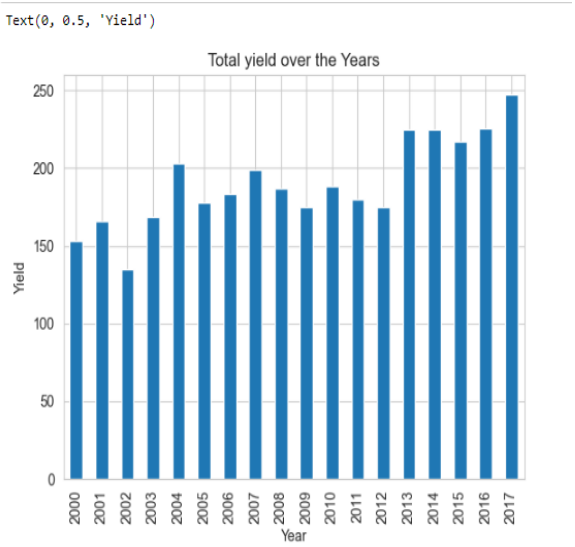
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 1, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 1, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=uint8)
```

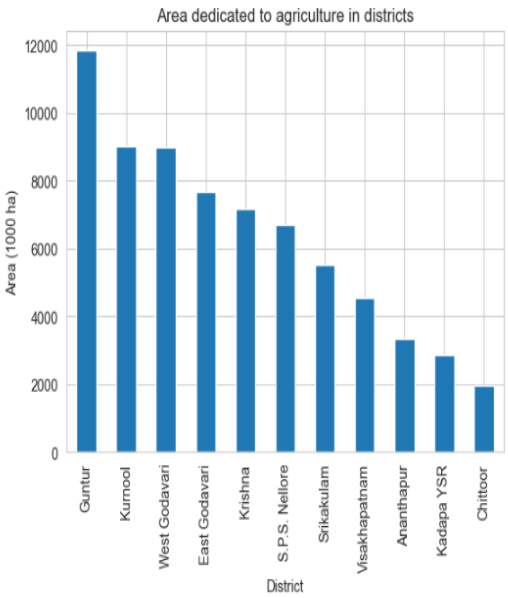
```
Accuracy score: 0.98
```

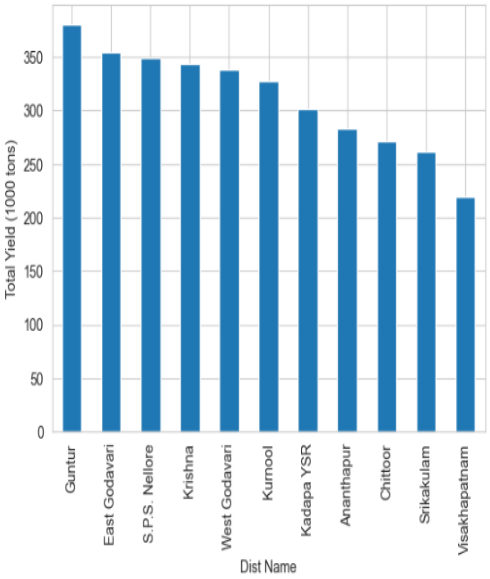
The accuracy score for the predictions made by the MultiOutputClassifier on the test set is 0.98, indicating a very high level of accuracy.

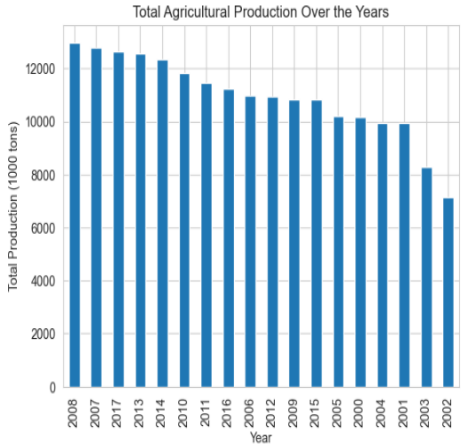
Crop Data Visualization







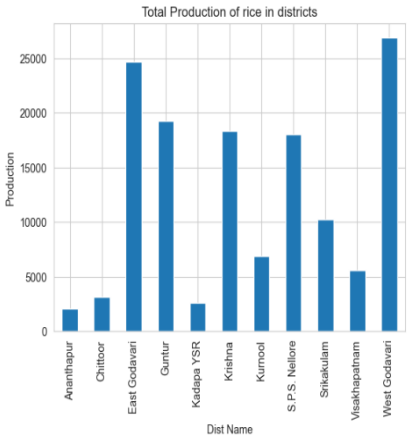


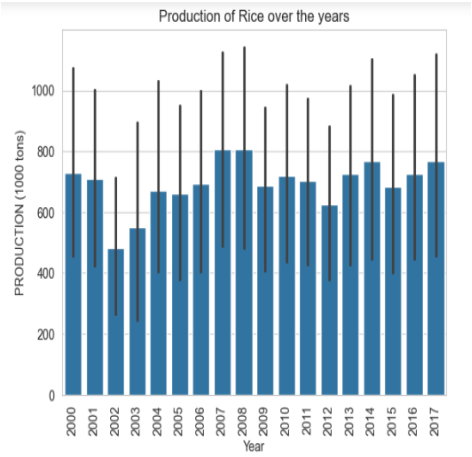


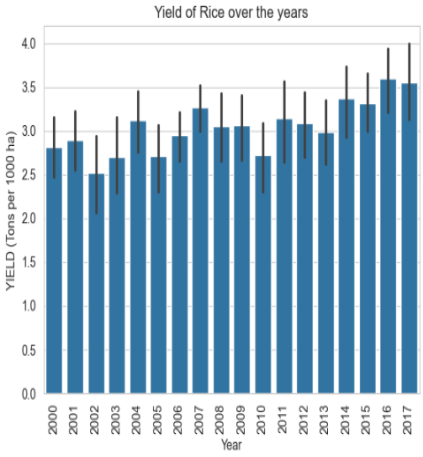
Observations:

- West Godavari district has maximum crop production
- Guntur district has maximum crop yield
- In 2008 crop Production was maximum
- Guntur district has maximum area dedicated to agriculture

Rice



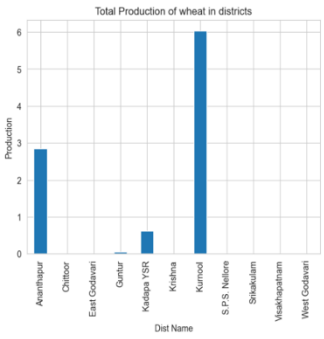


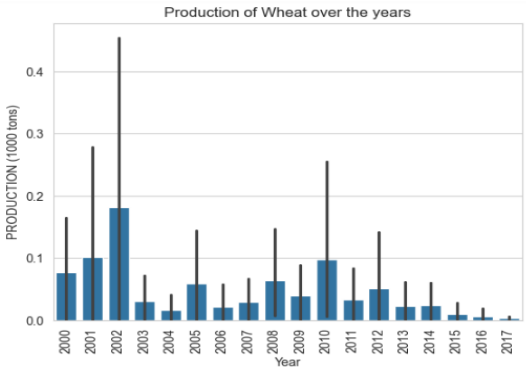


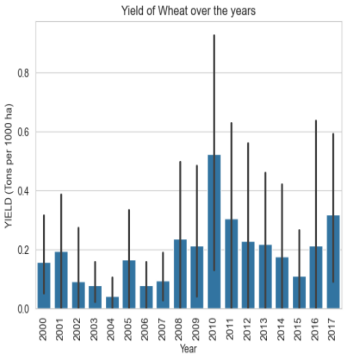
OBSERVATION:

- Rice Production is maximum in West Godavari district
- Production per unit Area of Rice was maximum in the year 2016 and 2017

WHEAT



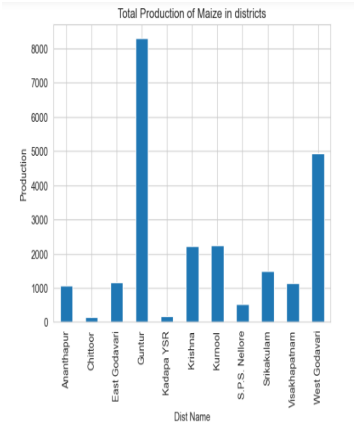


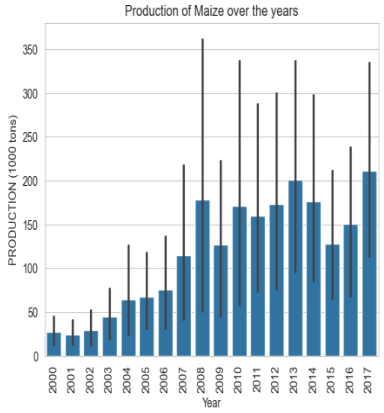


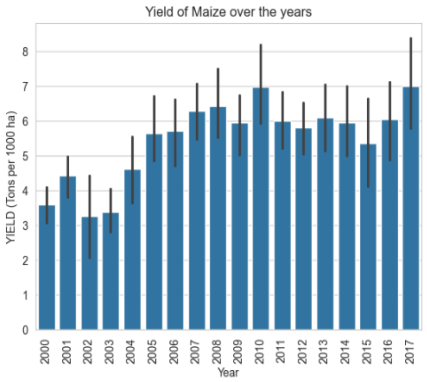
Observation

- Wheat Production is maximum in Kurnool district
- Production per unit Area of Wheat fluctuates over years, there was a sudden increase in the year 2010

MAIZE



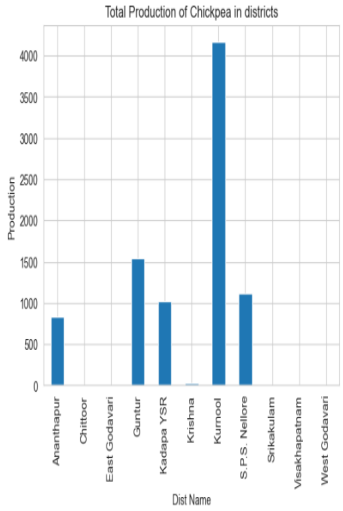


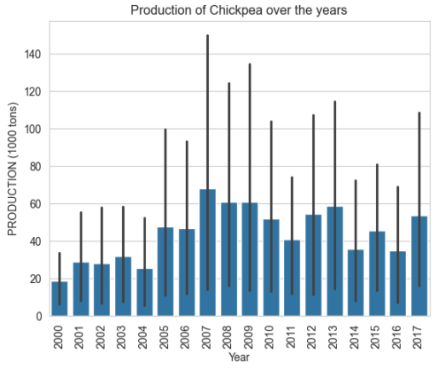


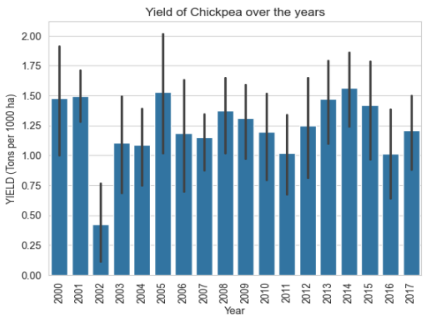
Observation

- Maize Production is maximum in Guntur district
- Production per unit Area of Maize increased after the year 2004 and was maximum in the year 2010 and 2017

CHICKPEA



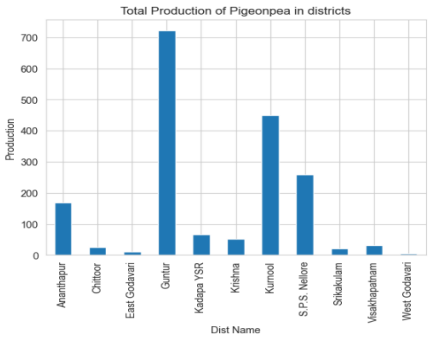


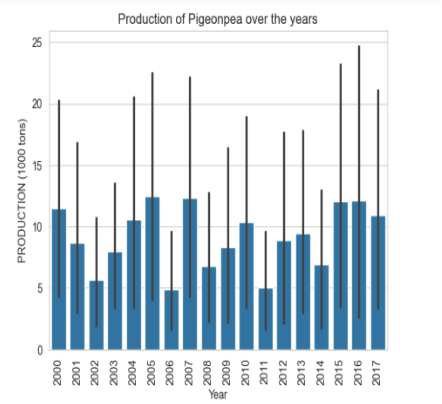


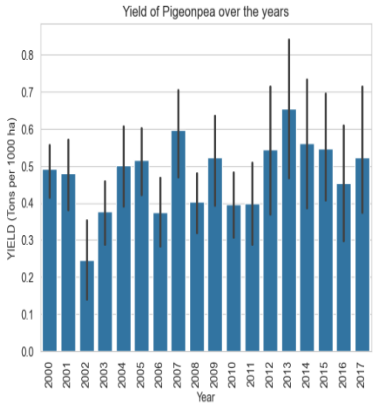
Observation:

- Chickpea Production is maximum in the Kurnool district
- Production per unit Area of Chickpea fluctuates over the years, sudden decrease in the year 2002 and maximum in the year 2005 and 2014

PIGEONPEA



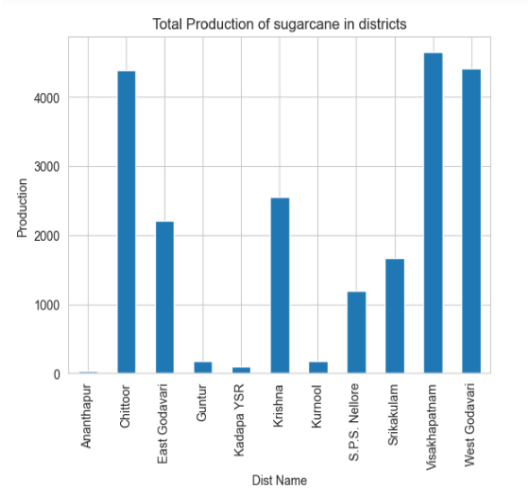


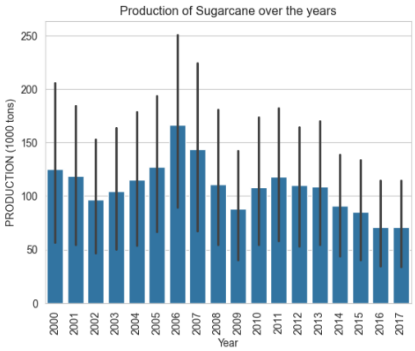


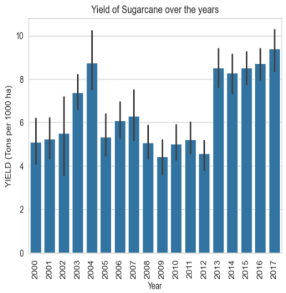
Observation

- Pigeonpea Production is maximum in Guntur District
- Production per unit Area of Pigeonpae fluctuates over the years and maximum in the year 2013

SUGARCANE



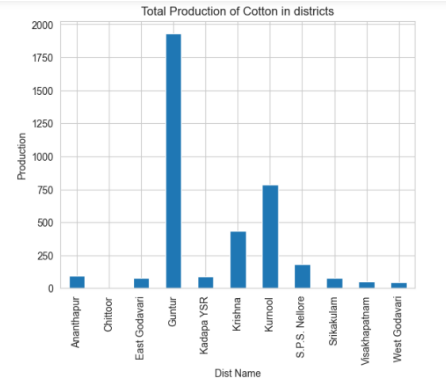


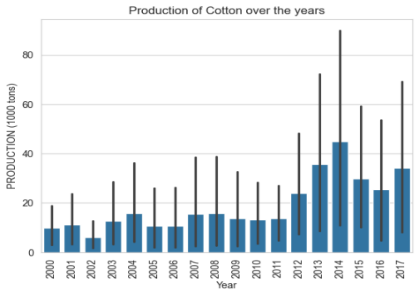


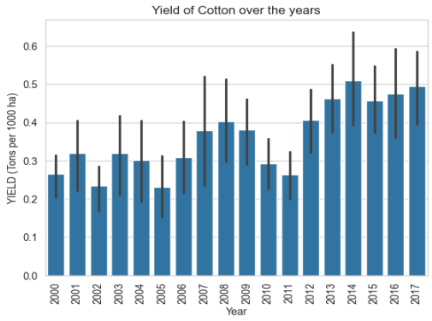
Observation:

- Sugarcane Production is maximum in Visakhapatnam district
- Production per unit Area of Sugarcane inreased from year 2013 after 2004

COTTON



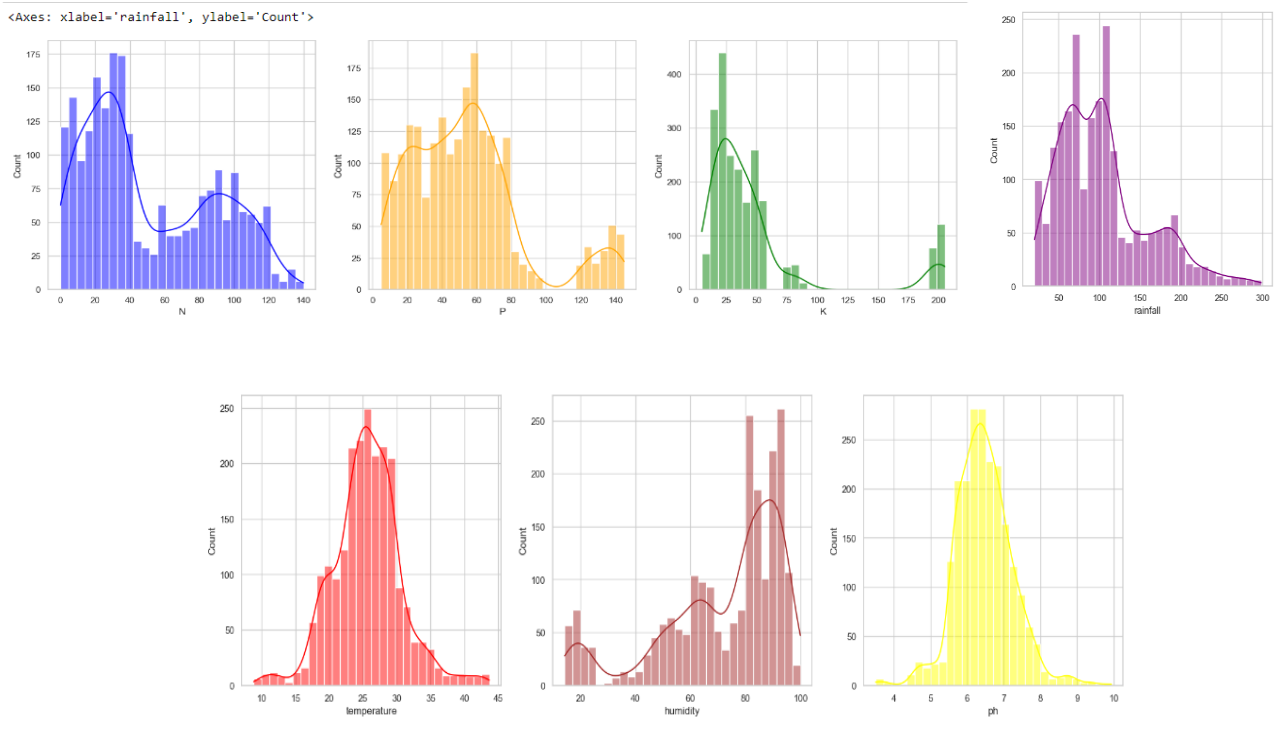




Observation:

- Cotton Production is maximum in Guntur district
- Production per unit Area of Cotton fluctuates over the years

Parameter Data Visualisation



Data Visualization in PowerBI

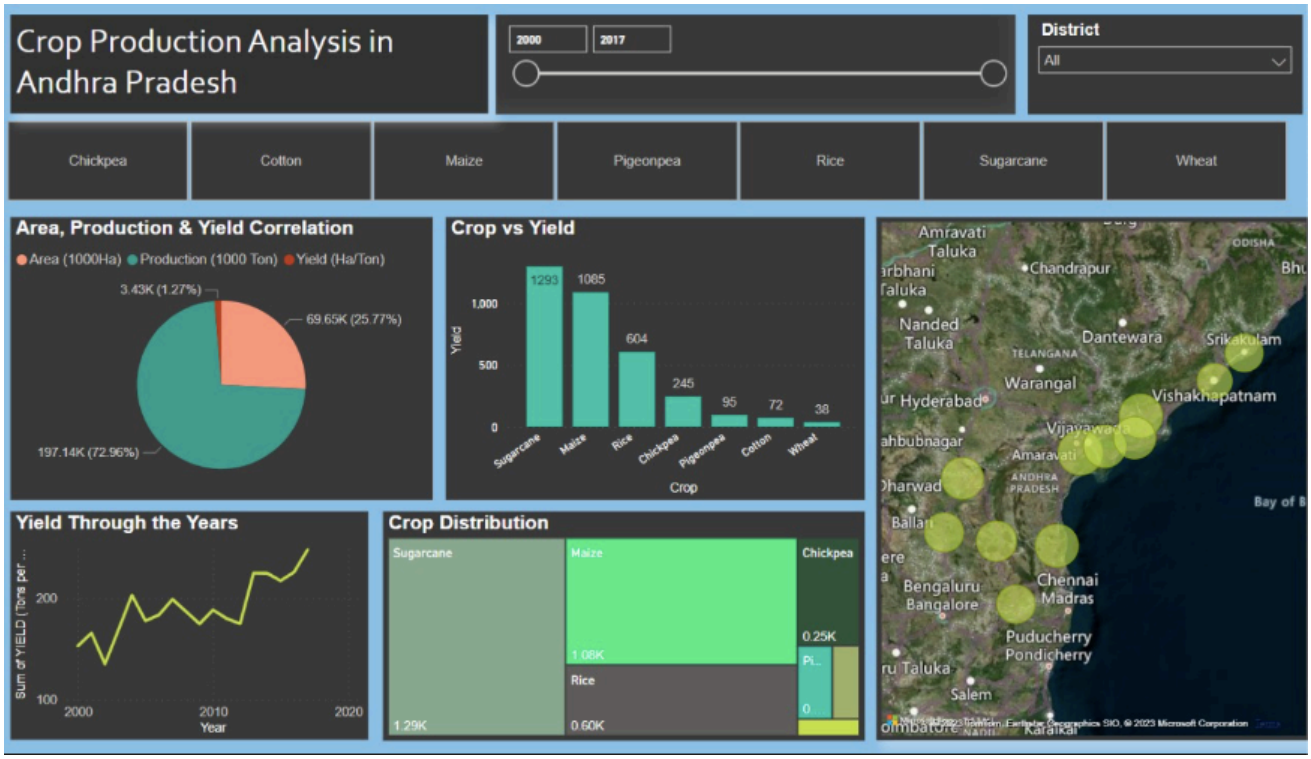


Figure B.1

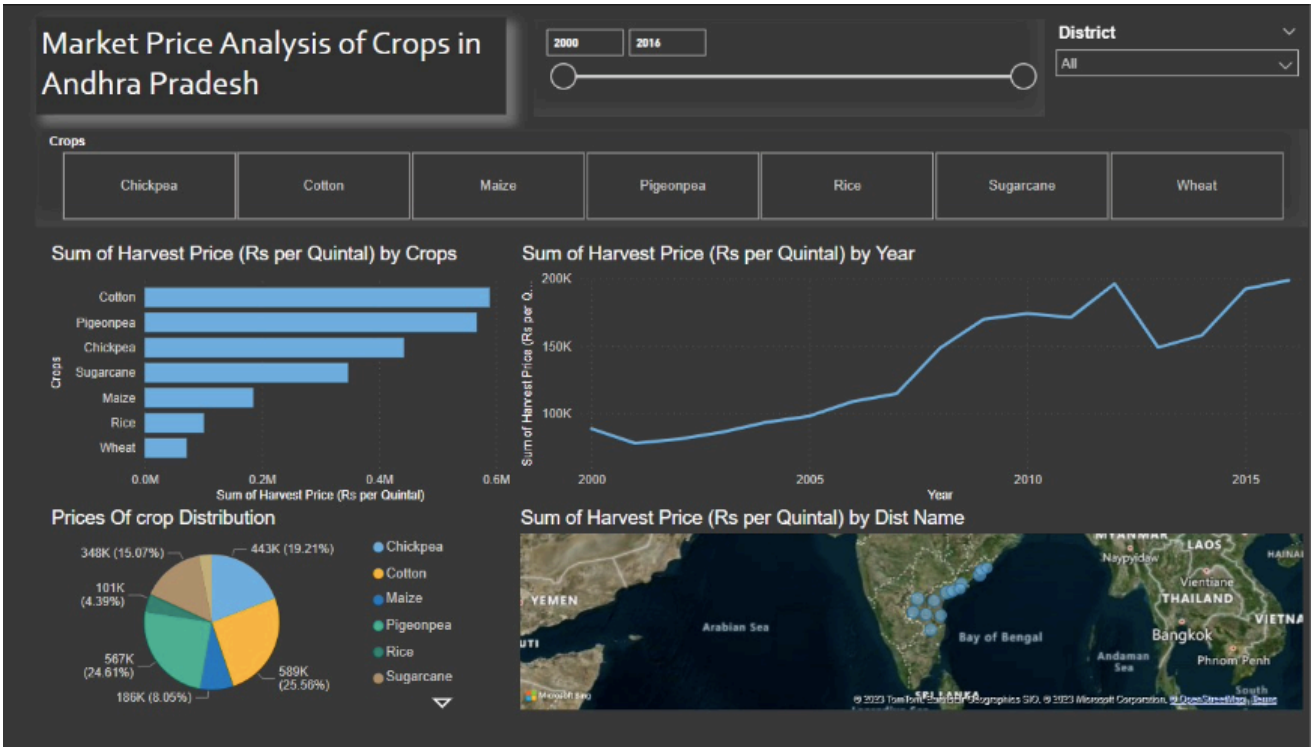


Figure B.2