First Year (Semester-2) Research Assignment on

Crop Recommendation System EDA

in partial fulfilment of the requirement for the successful completion of semester 2 of MSc Big Data Analytics

Submitted By

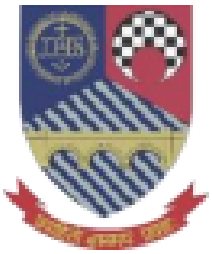
24-PBD-045

Krishna Saverdekar

(Semester – II MSc. BDA)

Under the supervision of

Prof. Hemal Desai



2023-2024

Department of Computer Sciences (MSc. BDA)

St. Xavier's College (Autonomous) Ahmedabad – 380009

DECLARATION

I, the undersigned solemnly declare that the research assignment *Crop Recommendation System EDA* is based on my work carried out during the course of our study under the supervision of *Prof. Hemal Desai*. I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that

• The work contained in the report is original and has been done by me under the general supervision of my supervisor.

• The work has not been submitted to any other Institution for any other degree / diploma / certificate in this university or any other University of India or abroad.

• We have followed the guidelines provided by the department in writing the report.

Krishna Saverdekar

24-PBD-045

MSc. BDA (Big Data Analytics)

St. Xavier's College (Autonomous), Ahmedabad

TABLE OF CONTENTS

Section No.	Title of section (Heading)	Page Number
1	Abstract	4
2	Introduction	4
3	Review of Literature	4
4	Objective, Data & Methodology	6
5	Data and Data Analysis	6
6	Findings and Conclusions	12
7	References	14

1. Abstract

This paper presents a study on crop recommendation by analysing soil properties and weather conditions using exploratory data analysis. The primary aim is to understand the relationships between various agricultural parameters—including soil nutrient levels, temperature and humidity—and their impact on crop selection. Using a dataset, correlation matrices, and visualizations (tables and charts) were generated to reveal key trends and insights. The findings highlight significant correlations among input features, which could serve as a basis for developing predictive models for crop recommendation. The study concludes by suggesting further work on integrating advanced machine learning algorithms to build robust crop prediction systems.

2. Introduction

Agriculture remains a cornerstone of global food security and economic development. With increasing climate variability and the need to optimize land use, predictive models that harness soil and weather data are critical. Crop prediction systems not only help farmers select the most suitable crops based on environmental conditions but also support decision-makers in resource allocation. This paper employs exploratory data analysis (EDA) to investigate how various factors such as soil nutrient content and weather conditions influence crop choices. By thoroughly understanding the underlying data patterns, this research lays the groundwork for developing accurate and reliable crop prediction models.

3. Literature Review

Several studies within the agricultural domain leverage data-driven approaches to provide insights for crop-related decisions, including recommendations. Machine learning techniques are increasingly being employed to predict crop production and recommend suitable crops based on various factors such as climate data and soil conditions.

The study by Banerjee et al. focuses on machine learning-based crop prediction using region-wise weather data. They mention that machine learning algorithms are used to

predict crop production based on climate data, which can benefit farmers in increasing crop production. Their proposed model aims to provide farmers with production estimates based on meteorological conditions and area under cultivation, helping them decide whether to produce a specific crop or choose an alternative if yield forecasts are negative.

Yadav et al. utilize EDA and various machine learning models to forecast crop yields using USDA data. Their research emphasizes identifying the underlying factors that affect yield through thorough EDA, encompassing weather patterns and USDA-sourced soil composition. This EDA helps gain crucial insights into the variables impacting crop output differences, which then forms the basis for their machine learning modeling aimed at optimizing crop production estimates. The ultimate goal is to support stakeholders in making well-informed agricultural decisions and enabling sustainable practices.

Bhuyan et al. specifically addresses crop type prediction using statistical analysis and machine learning. Their study provides a statistical look at features like nitrogen, phosphorus, potassium, pH, temperature, humidity, and rainfall to indicate the best crop type in an Indian smart city context. They perform an in-depth statistical analysis of these soil and climatic attributes to understand their influence on crop selection. This statistical analysis, which falls under EDA, helps in understanding the relationships and distributions of the data before applying machine learning algorithms like k-NN, SVM, RF, and GB trees to model and predict the crop type. The study highlights the importance of understanding how attributes like humidity and rainfall in different Indian locations lead to specific crop cultivation, as well as the role of potassium and nitrogen content in soil for crop selection. They even identify 'rainfall' as the most important attribute using a feature selection algorithm, which is a step often informed by initial EDA. Their work aims to aid farmers in selecting the best crop for their region's soil and climate by providing a system that takes soil characteristics and climate as input and recommends the most suitable crop.

Jabed and Murad's review also highlights the significance of environmental and agricultural data, including temperature, rainfall, soil type, and humidity, for accurate crop yield estimation, which is related to crop recommendation as understanding potential yield is crucial for making informed planting decisions. They note that various

classification and regression techniques have been successful in agricultural production prediction. Their systematic literature review explores the features utilized for crop yield prediction through machine learning and deep learning, including soil type, temperature, humidity, and rainfall, which are also critical factors in determining suitable crop types. They also discuss the increasing use of remote sensing for data collection on environmental conditions and crop growth, providing valuable data for EDA and subsequent modeling for crop-related recommendations. The review emphasizes the need for a deeper insight into the variables and factors influencing crop production predictions, suggesting that while many features are used, further exploration is needed to identify those with the most significant impact, a process often initiated with EDA.

4. Objective

- To perform exploratory data analysis (EDA) on a dataset that includes soil properties and weather conditions.
- To identify key relationships among soil nutrient levels, weather parameters, and crop types.
- To generate visual insights (tables and charts) that highlight trends and correlations within the dataset.
- To propose directions for further development of predictive models based on the insights derived from the EDA.

5. Data & Methodology

The dataset is comprehensive, encompassing various key factors critical to machine learning-based crop recommendation systems. The soil properties dataset includes detailed information such as specific locations identified by latitude and longitude coordinates, soil pH, soil color, surface soil composition, electrical conductivity, and a range of soil macro and micronutrients. These factors are essential in determining the suitability of different crops to various soil types. The crop type information in the dataset primarily comprises cereals, reflecting the major crops under study.

The climate features dataset includes a broad spectrum of environmental conditions crucial for crop growth. This includes specific location data (latitude and longitude coordinates), as well as seasonal variations such as maximum and minimum temperatures, precipitation levels, humidity, wind speed and direction, surface pressure readings, and cloud cover assessments. These features provide insights into the climatic conditions that crops will experience throughout the growing season. The climate data has been sourced from the National Aeronautics and Space Administration's (NASA) cloud infrastructure.

6. Data and Data Analysis

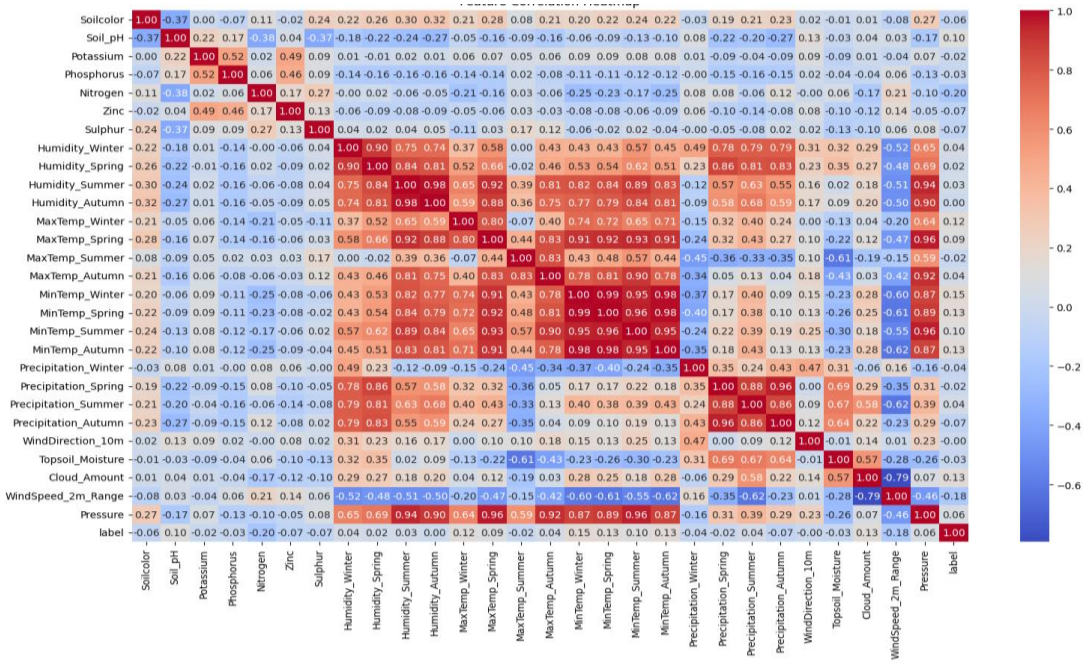


Fig 6.1. Correlation Heatmap

Most features have weak correlation with crop type or yield. The highest correlation appears in pressure (0.06) and humidity (0.07), but these values are still low, meaning no single environmental factor directly determines crop type alone. This suggests that crop yield or selection depends on multiple factors together rather than just one.

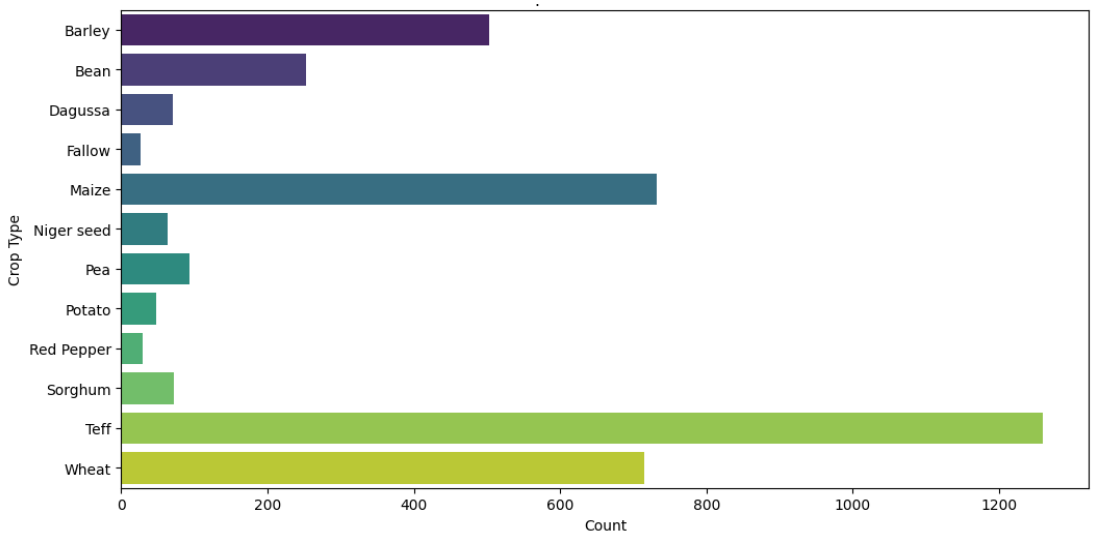


Fig. 6.2. Crop Label Distribution

Teff has the highest count, significantly more than any other crop. Maize and Wheat also have high representation in the dataset. Barley and Bean have moderate counts compared to the top crops. Degussa, Niger Seed, Pea, Potato, and Red Pepper have much lower counts.Fallow land (unused land) is present but in small quantities.

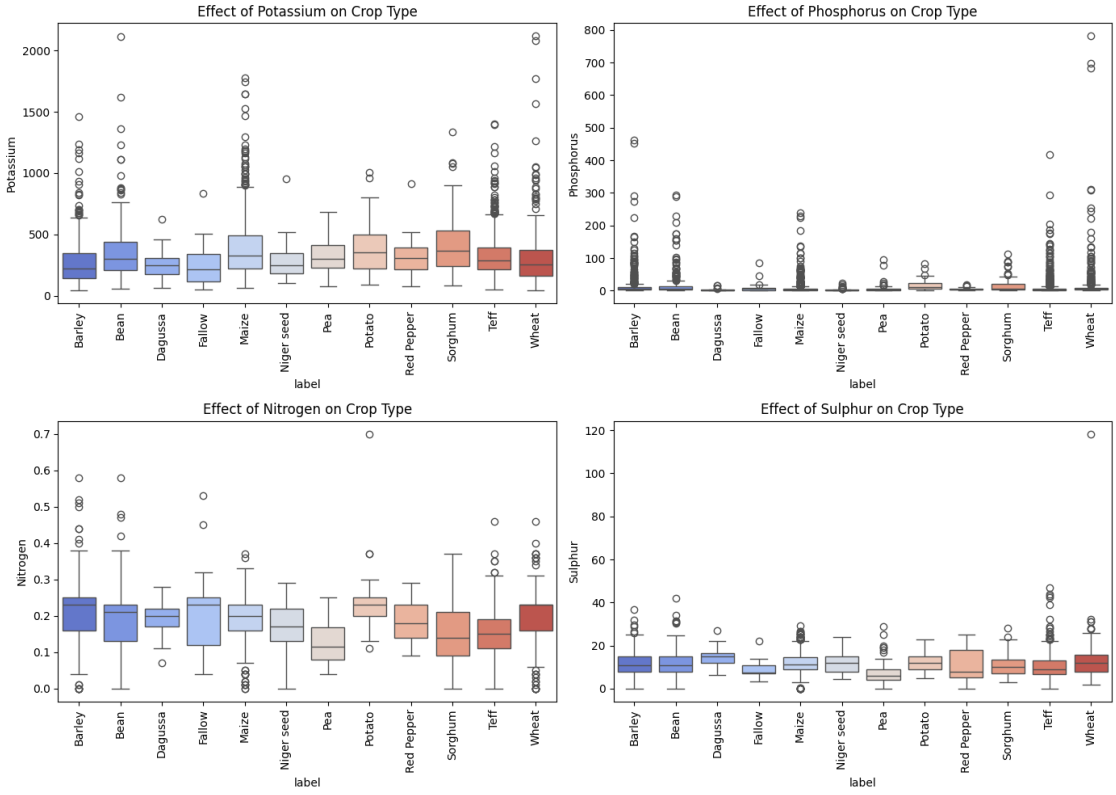


Fig.6.3. Effect of Soil nutrients across different crop types

1. Effect of Potassium on Crop Type (Top-left)

Potassium levels vary significantly across crops. Outliers: Many extreme outliers, indicating that some crops can thrive even at very high Potassium levels. Crop-Specific Observations: Teff & Wheat: Have higher Potassium concentrations. Red Pepper & Sorghum: Show moderate levels of Potassium. Barley, Beans, and Peas: Have lower median Potassium values.

2. Effect of Phosphorus on Crop Type (Top-right)

Phosphorus levels are generally low across most crops. Outliers: Many extreme values, especially for Teff, Wheat, and Sorghum. Crop-Specific Observations: Barley & Beans: Have very low Phosphorus levels. Teff & Wheat: Have a wider range of phosphorus levels, possibly indicating higher adaptability.

3. Effect of Nitrogen on Crop Type (Bottom-left)

Nitrogen levels are relatively low compared to Potassium. Outliers: Some crops show very high Nitrogen levels.

Crop-Specific Observations: Barley, Beans & Fallow: Have slightly higher Nitrogen levels. Teff & Wheat: Have a broader range of Nitrogen levels.

4. Effect of Sulphur on Crop Type (Bottom-right)

Sulphur levels are mostly low across crops. Outliers: A few extreme high values in Wheat and Red Pepper. Crop-Specific Observations: Barley & Beans: Have the lowest Sulphur levels. Red Pepper & Sorghum: Show moderate Sulphur concentrations. Wheat & Teff: Have the highest Sulphur levels.

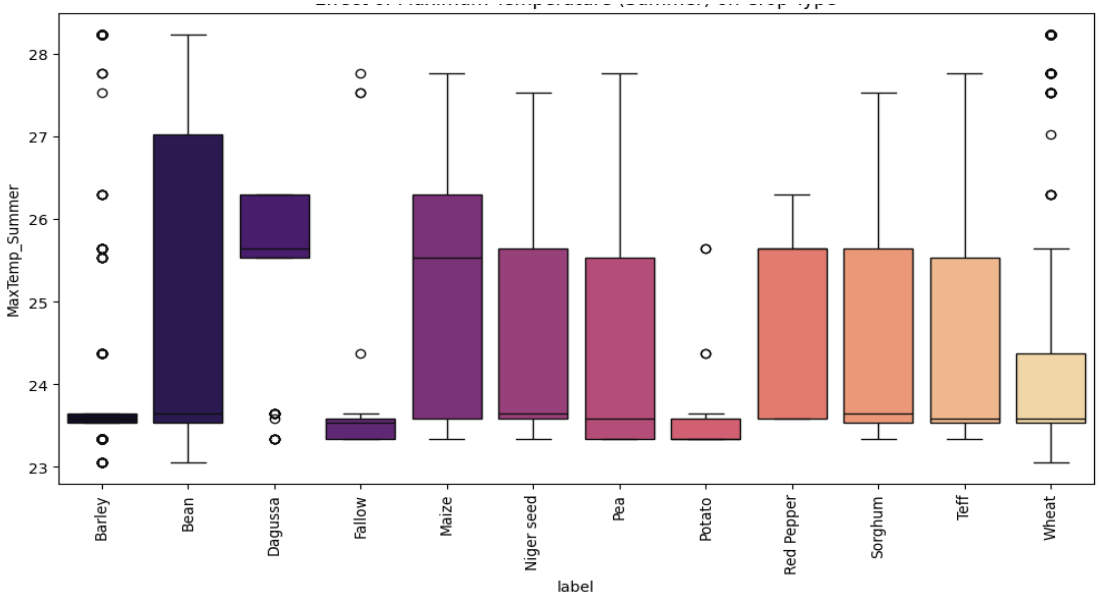


Fig.6.4. Effect of Maximum Temperature on Crop Type

Some crops have a narrow range, meaning they thrive in specific temperature conditions. Others have a wide range, suggesting they can tolerate a broad spectrum of temperatures. Beans, Peas, and Maize show higher temperature variability, with some instances reaching up to 27–28°C.Sorghum, Teff, and Red Pepper also show higher maximum temperatures. These crops may be more heat-resistant or require warmer climates to grow. Barley, Potato, and Fallow crops have a narrow temperature range (~23-24°C). This suggests they are more temperature-sensitive and require stable conditions. Potato and Barley especially show low median temperatures, meaning they are suited for cooler climates. Outliers Indicating Extreme Cases: Some crops (e.g., Beans, Wheat, Sorghum) have outliers reaching above 28°C, indicating that certain cases of these crops can survive extreme heat. On the other hand, Barley and Potato have very few outliers, showing they are not well adapted to extreme heat.

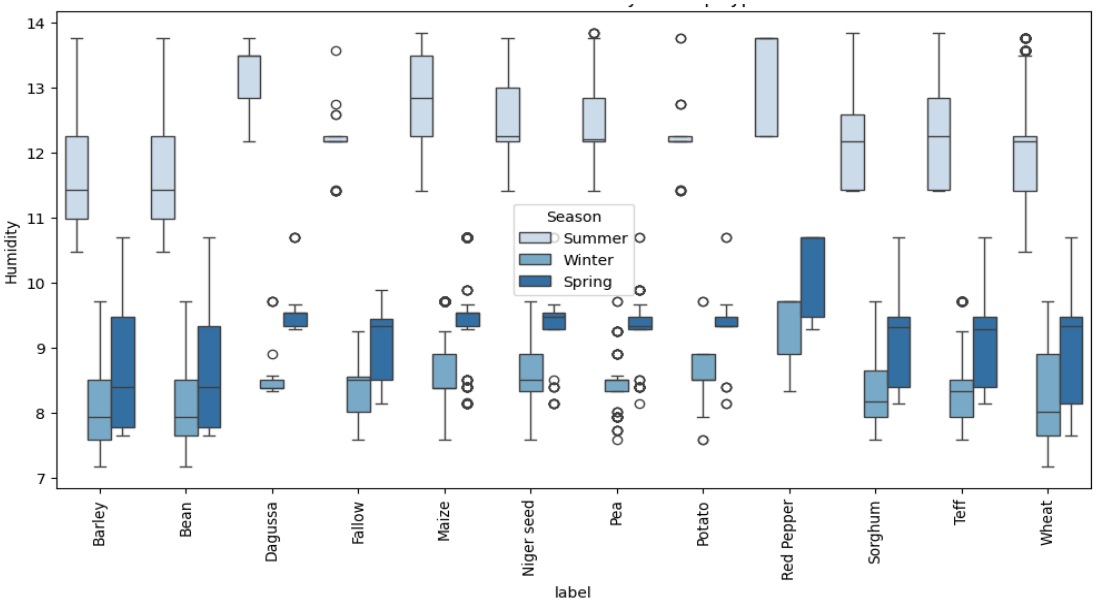


Fig.6.5. Effect of Seasonal Humidity on Crop Type

Barley, Bean, and Teff experience higher humidity levels in Summer, showing that they might thrive in more humid conditions. Degussa, Pea, and Potato show consistent humidity levels across all seasons, meaning they are more stable in varying climates. Red Pepper and Fallow crops have lower humidity tolerance, with Spring and Winter showing lower humidity values. Some crops (e.g., Niger Seed, Wheat, and Maize) have higher variability in humidity across seasons, indicating they can adapt to a range of moisture conditions. Degussa and Potato have very low humidity variations, suggesting they require a stable humidity environment.

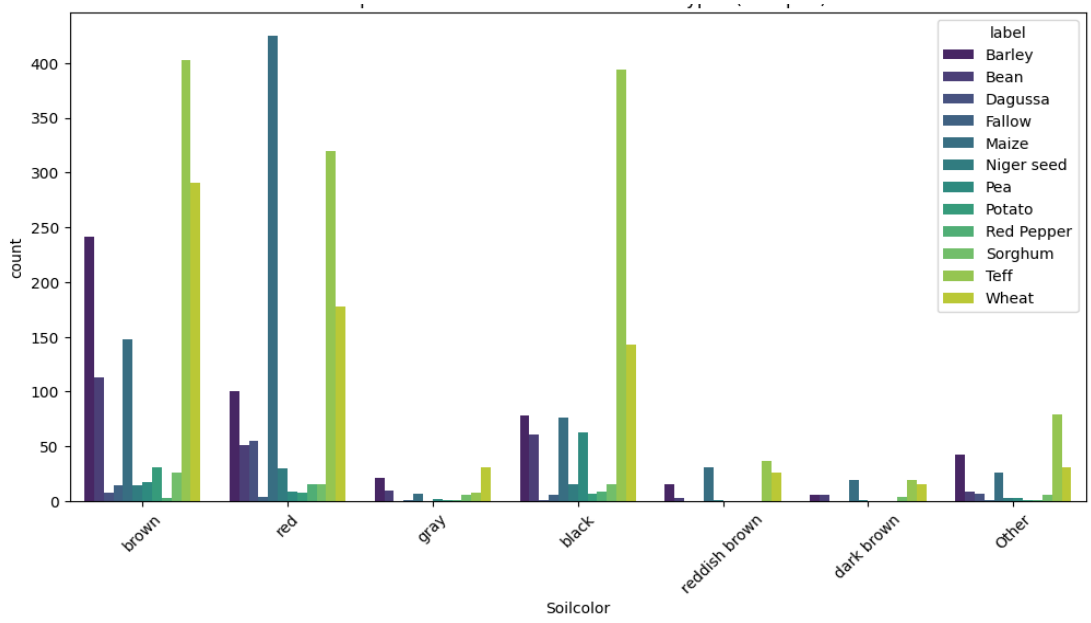


Fig.6.6 Crop Distribution Across Different Soil Types

Brown, Red, and Gray soils are the most prevalent. Black and Dark Brown soils also have moderate representation. Other soil types (light brown, dark grey, reddish brown, etc.) have fewer crops. Wheat, Teff, and Sorghum are mostly found in Brown, Red, and Gray soils. Barley, Bean, and Maize show diversity in multiple soil colors, but Brown and Red remain dominant. Degussa and Niger seed prefer Gray and Black soils. Potato and Pea have lower counts across all soil types, indicating a specific soil preference. Brown and Red soils seem to support a higher variety of crops, making them fertile and well-balanced. Black soil appears suitable for specific crops like Maize, Barley, and Niger Seed. Light Red, Dark Gray, and Reddish-Brown Moist soils have very few crops, indicating low fertility or limited use in agriculture.

7. Key Findings and Conclusion

The analysis of soil nutrients and environmental factors across different crops highlights key insights that can aid in optimizing agricultural productivity. One major finding is that nutrients are unevenly distributed across crops, meaning different crops require varying concentrations of essential elements like Potassium, Phosphorus, and Sulphur. Crops like Teff and Wheat demonstrate higher nutrient tolerance, whereas Beans and Barley thrive in lower nutrient conditions. Additionally, outliers in nutrient distribution suggest that certain crops have unique adaptability, allowing them to grow even in extreme soil conditions.

From a climate perspective, temperature and humidity play a crucial role in crop selection. Heat-resistant crops such as Beans, Peas, Sorghum, and Teff can tolerate higher summer temperatures, whereas cool-climate crops like Barley and Potato prefer lower temperature ranges. Notably, crops like Bean and Teff require high humidity in summer, making them ideal for humid climates, while Maize and Wheat exhibit wide adaptability, allowing them to thrive in various seasonal conditions. This knowledge can help farmers decide the most suitable planting seasons based on climate preferences, reducing risks associated with unpredictable weather.

The study also reveals that soil type significantly impacts crop selection. Brown and Red soils emerge as the most versatile, supporting multiple crops, whereas Gray and Black soils are more specialized, favoring crops like Degussa and Niger Seed. Some soils, such as Light Red, Dark Gray, and Reddish Brown Moist, support fewer crops, indicating lower fertility levels. By understanding these soil-crop relationships, farmers can make informed decisions about soil suitability rather than forcing unsuitable crops, leading to better yields and sustainable farming.

To enhance productivity, several actionable steps can be implemented. Conducting soil testing before cultivation allows farmers to determine nutrient composition and apply customized fertilizers to balance soil nutrients effectively. Over-fertilization, particularly with Phosphorus and Sulphur, should be avoided to prevent reduced productivity.

Conclusion

This data-driven approach to farming provides valuable insights into soil suitability, seasonal impacts on crop yield, and the importance of balanced soil nutrients. By implementing sustainable techniques like crop rotation, irrigation management, and soil testing, farmers can optimize yields while preserving soil health. The findings emphasize the significance of precision agriculture, where decisions are based on data analysis rather than traditional farming practices, ensuring higher productivity, better resource utilization, and long-term sustainability.

8. References

Alemu, S. (2024). *Crop recommendation using soil properties and weather prediction dataset* [Dataset]. Mendeley Data. https://doi.org/10.17632/8V757RR4ST.1

Banerjee, S., Chakraborty, S., & Mondal, A. C. (2023). Machine learning based crop prediction on region wise weather data. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(1), 145–153. https://doi.org/10.17762/ijritcc.v11i1.6084

Bhuyan, B. P., Tomar, R., Singh, T. P., & Cherif, A. R. (2022). Crop type prediction: A statistical and machine learning approach. *Sustainability*, *15*(1), 481. https://doi.org/10.3390/su15010481

Jabed, Md. A., & Azmi Murad, M. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, *10*(24), e40836. https://doi.org/10.1016/j.heliyon.2024.e40836

Optimizing crop yield prediction: Data-driven analysis and machine learning modeling using usda datasets – current agriculture research journal. (2024, April 15). https://www.agriculturejournal.org/volume12number1/optimizing-crop-yield-prediction-data-driven-analysis-and-machine-learning-modeling-using-usda-datasets/