EXPLORATORY DATA ANALYSIS (EDA) USING PYTHON

CROP RECOMMENDATION SYSTEM USING SOIL AND CLIMATE DATA

Name: Aishwarya Gunaseelan
Domain: Data Science

1. Introduction

Agriculture is one of the most important sectors of the economy, and crop productivity highly depends on soil properties and climatic conditions. Farmers often face difficulty in selecting the most suitable crop due to lack of scientific analysis of soil nutrients and weather parameters.

This project focuses on performing **Exploratory Data Analysis (EDA)** on a crop recommendation dataset obtained from Kaggle. The dataset includes soil nutrient values and climatic factors such as temperature, humidity, pH, and rainfall, along with the recommended crop. EDA helps in understanding the data distribution, identifying relationships between variables, and extracting meaningful insights for agricultural decision-making.

2. Aim / Objective

The main objectives of this project are:

- To understand the structure and characteristics of the crop recommendation dataset
- To analyse soil nutrients and climate parameters
- To visualize the distribution of features using graphs
- To identify relationships between environmental factors and crops
- To derive useful insights that support crop selection decisions

3. Business Problem

Farmers often rely on traditional knowledge to decide which crop to cultivate. However, incorrect crop selection may result in low yield, financial loss, and inefficient use of soil nutrients.

The business problem addressed in this project is **how to recommend suitable crops based on soil and climatic conditions**. By analysing historical data, this project helps understand which crops perform well under specific environmental conditions, thereby improving productivity and sustainability in agriculture.

4. Project Workflow

The following steps were followed in this project:

1. Selection of crop recommendation dataset from Kaggle
2. Importing required Python libraries
3. Uploading and loading the dataset in Google Colab
4. Understanding dataset features and structure
5. Data cleaning and validation
6. Univariate analysis using histograms
7. Bivariate analysis using boxplots
8. Multivariate analysis using correlation heatmap
9. Interpretation of results and insights
10. Documentation and conclusion

5. Data Understanding

The dataset contains **2200 rows and 8 columns**, described below:

Feature	**Description**
N	Nitrogen content in soil
P	Phosphorus content in soil
K	Potassium content in soil
temperature	Temperature in °C
humidity	Relative humidity (%)
ph	Soil pH value
rainfall	Rainfall in mm
label	Recommended crop

Observations:

- All input features are numerical
- The target variable (label) is categorical
- No missing values were found
- Data is well structured and clean

6. Data Cleaning

The following data cleaning steps were performed:

- Checked for missing values using df.isnull().sum()
- Verified duplicate records using df.duplicated()
- Removed duplicate rows (if present)
- Ensured correct data types for all columns

The dataset was found to be clean and required minimal preprocessing.

7. Derived Metrics

A new derived metric called **Nutrient Index** was created to represent average soil fertility:

$$\text{Nutrient Index} = \frac{N + P + K}{3}$$

This metric helps summarize overall soil nutrient strength in a single value.

8. Filtering Data

Data filtering was applied to analyse specific conditions such as:

- Crops with high rainfall requirements
- Crops grown in acidic or neutral pH soils
- Soil conditions with high nitrogen or potassium content

Filtering allows focused analysis of specific agricultural scenarios.

9. Statistical Analysis

Correlation analysis was performed to study relationships between numerical variables.

Key Findings:

- Soil nutrients (N, P, K) show low correlation with climatic variables
- Temperature and humidity show moderate correlation
- Rainfall varies significantly across crop types

A correlation heatmap was used for better visualization and interpretation.

10. Exploratory Data Analysis (EDA)

10.1 Univariate Analysis

- Histograms were plotted for soil nutrients and climate variables
- Crop frequency distribution was visualized using count plots

10.2 Bivariate Analysis

- Boxplots were used to analyse rainfall vs crop
- Temperature variation across crops was visualized

10.3 Multivariate Analysis

- Correlation heatmap displayed relationships among multiple variables
- Helped identify influential factors for crop selection

11. Insights

1. Different crops require different combinations of soil nutrients and climate conditions
2. Most crops prefer near-neutral pH values
3. Rainfall plays a major role in determining crop suitability
4. Temperature and humidity vary significantly across crops

5. Nutrient-rich soils support crops such as rice, maize, and sugarcane
6. Correlation analysis confirms independence of many environmental variables

12. Conclusion

This project successfully applied Exploratory Data Analysis techniques to analyse the crop recommendation dataset. The study revealed important patterns between soil nutrients, climatic factors, and crop selection. The visualizations and statistical summaries provided meaningful insights that can help farmers and agricultural planners make informed decisions.

The project demonstrates how data science techniques can be effectively applied in agriculture to improve productivity and promote sustainable farming practices.