Data Analytics and Visualization Project

Crop Recommendation and Visualization

Problem Statement:

- To design a data-driven crop recommendation system that predicts optimal crops based on nutrient and climate data by training a machine learning model .
- Visualize critical agricultural insights through interactive dashboards.
- Exploratory visualizations of plant health from image datasets .

Problem Solved :

- Empowers farmers and policymakers to make informed, data-backed decisions for improved agricultural productivity.

Dataset Specification Statement:

This project utilizes 15 crops and its related

- CSV datasets containing [N, P, K levels, pH, temperature, humidity, rainfall] for crop recommendation modeling
- Plant image datasets (approx 450 diseased and 50 healthy for each crop) for exploratory visualization.

Workflow

01

Data Acquisition & Preprocessing : CSV + images dataset

02

Training Crop Recommendation Model and ANOVA testing

03

Visualisations of the images as well as the CSV dataset

04

Interactive Streamlit Dashboard

Crop Recommendation Model

Crop Recommendation Model

- Goal: Predict the most suitable crop based on soil and climate data.
- Model Used: Random Forest Classifier.
- Dataset Attributes: Includes N, P, K, temperature, humidity, pH, and rainfall.

Data Preprocessing

- Fixed column name inconsistencies.
- Filled missing values using median crop-wise values.
- Removed duplicate entries.
- Encoded crop labels using LabelEncoder.

Evaluation Metrics

- Accuracy: 0.9967
- Precision: 0.9968
- Recall: 0.9967
- F1 Score: 0.9967

CSV Visualization

Crop Distribution Analysis

A horizontal bar plot showing the number of samples for each crop in the dataset.
Each crop has exactly 100 samples, ensuring a balanced dataset .

Feature Correlation Heatmap

A heatmap showing the pairwise correlation between all numeric features (e.g., N, P, K, temperature, humidity, pH, rainfall).
Helps in identifying strong relationships or redundant features in the dataset.

Histogram with KDE Curve

Histograms with KDE curves for each feature (N, P, K, temperature, humidity, pH, rainfall) show how values are distributed across the dataset.
Used to understand data spread, central tendency, skewness, and presence of outliers.

Feature Pair Relationships (Scatter Matrix)

A scatter matrix showing pairwise relationships between all numerical features (N, P, K, temperature, humidity, pH, rainfall).
Diagonal plots show KDE (distribution) for each feature.
Used to visually detect correlations, trends, and clusters among variables.

3D Scatter Plot

A 3D scatter plot visualizing the distribution of Nitrogen (N), Phosphorus (P), and Potassium (K) requirements across all crop types.
Each point represents a crop sample, color-coded by crop label, helping identify nutrient demand patterns and clusters in NPK space.

Average Nutrient Levels Heatmap

This heatmap displays the average values of Nitrogen (N), Phosphorus (P), and Potassium (K) required by each crop.
Color intensity indicates nutrient concentration—darker shades represent higher values.

Nutrient and Climate Feature Distribution Boxplot

This set of boxplots illustrates the distribution of key features (Nitrogen, Phosphorus, Potassium, Temperature, Humidity, pH, and Rainfall) for each crop type. The plots display how these variables vary across different crops, highlighting median values, quartiles, and potential outliers.

Images Visualisation

Class Distribution Histogram	Bar plot showing image count per plant class, highlighting data balance/imbalance across categories. Labels display exact counts for quantitative comparison
Comparative Leaf Sample Gallery	Side-by-side display of healthy vs. diseased leaf samples for visual quality assessment and dataset familiarization. Highlights morphological differences between conditions.
HSV Hue Channel Distribution Analysis	Comparative histograms of dominant hue values (0-1 scale) showing color shifts between healthy and diseased leaves across multiple crop types, with statistical significance indicators.

Disease Anomaly Heatmaps

Comparative visualization of healthy vs. diseased leaves with heatmap overlays highlighting regions of significant hue variation, quantifying disease spread patterns and severity.

Normal Hue Distribution Density Analysis

Kernel Density Estimation (KDE) plots comparing healthy (green) vs diseased (red) leaf hue distributions, revealing disease-induced color shifts through probability density curves.

Future Enhancements

01

Model Improvement & Evaluation

- Hyperparameter Tuning: Fine-tune model's settings (like learning rate or tree depth) using techniques like Grid Search or Random Search to improve performance.
- Cross-Validation: Split your data into multiple subsets to train and validate the model several times.

02

Web Application for User Interaction

- Deploy the Model as a Web App: Build a user-friendly interface (using Flask or Streamlit) where users can enter data (like weather conditions or soil properties) and get predictions, making the model more accessible.

03

Model Monitoring & Retraining Pipeline

- Real-time Monitoring: Set up real-time monitoring of your model's performance (e.g., using MLflow or TensorBoard). This allows you to track any model drift and adjust the model over time.

04

Geospatial Visualizations for Crop Prediction

- Integrate Geospatial Data: Use geospatial data (e.g., GPS coordinates, satellite imagery) to visualize crop distributions and nutrient needs across different regions. This can enhance the model's predictions by incorporating location-based factors into crop recommendation systems.

Thank You!