Agricultural Data Analysis and Yield Prediction

**1. Introduction**

This project aims to analyze agricultural production data (Crops and Livestock) and develop a predictive model to estimate future production values. The goal is to extract meaningful insights and trends from historical data and apply machine learning to make data-driven forecasts.

**2. Approach Overview**

**Step 1: Data Acquisition & Preparation**

* Imported a multi-sheet Excel file containing agricultural statistics.
* Separated the data into two primary categories:
  + **Crops**
  + **Livestock**
* Created pivot tables for both datasets with:
  + Rows: Country, Item, Year
  + Columns: Element (e.g., Production, Yield)
  + Values: Corresponding numeric values
* This helped consolidate information like production and yield under a unified structure.

**Step 2: Data Cleaning & Encoding**

* **Missing Values:** Filled using forward fill, backward fill, or mean imputation depending on the nature of the feature.
* **Categorical Encoding:** Applied Label Encoding for Country and Item columns to prepare data for modeling.
* Saved the cleaned and processed datasets as CSV files for traceability.

**Step 3: Exploratory Data Analysis (EDA)**

* Used matplotlib and seaborn for visual exploration.
* Aggregated data to observe macro trends:
  + Total production per year (global level)
  + Production by country and by commodity
  + Yield comparison across years

**Step 4: Machine Learning Model**

* Chose **RandomForestRegressor** due to its robustness with non-linear data and its ability to model feature interactions.
* Feature set included encoded variables and numeric features from the pivoted data.
* Evaluated model using R², MAE, and RMSE.

**3. Visualizations & Analysis**

**A. Global Production Over Time**

**Visualization:** Line plot of total production (summed across countries) by year  
**Insight:**

* A gradual increase in total crop production was observed over the years.
* Sudden dips may correspond to known global events like droughts or pandemics.

**B. Top Producing Countries**

**Visualization:** Bar chart of top 10 countries by average production  
**Insight:**

* Certain countries consistently lead in agricultural output (e.g., India, USA, China).
* Insights can guide investment or policy targeting high-performing regions.

**C. Yield vs. Production**

**Visualization:** Scatter plot of Yield vs. Production  
**Insight:**

* A strong positive correlation in some crops suggests better farming practices.
* Outliers may represent either exceptionally productive areas or data entry issues.

**D. Commodity-Wise Production**

**Visualization:** Grouped bar chart for selected items (e.g., wheat, rice, maize, poultry)  
**Insight:**

* Highlights dominant crops in global supply.
* Identifies underperforming commodities which may be targeted for optimization.

**4. Key Findings**

* **Production Trends:** Global agricultural production has shown a steady rise, but growth is uneven across regions.
* **Regional Insights:** Some countries dominate specific commodities (e.g., Brazil in livestock, India in cereals).
* **Feature Importance:** In the Random Forest model, the most important features were:
  + Item type (crop/livestock)
  + Year
  + Area (country)
  + Historical yield and production values
* **Predictive Modeling:** The RandomForestRegressor achieved satisfactory accuracy in predicting production with relatively low error, making it suitable for forecasting purposes.

**5. Actionable Insights**

* **Policy Implications:** Countries with low yield but large production areas could benefit from improved agricultural practices or technology adoption.
* **Investment Opportunities:** Identify high-growth commodities for strategic investments or subsidies.
* **Forecasting Use:** The model can be integrated into dashboards for annual production forecasting to help with supply chain planning.

**6. Conclusion**

The project successfully processed, analyzed, and modeled complex agricultural data to derive valuable insights and build a predictive system. The structured pipeline—from raw Excel to trained ML model—can serve as a template for similar agricultural analytics tasks or real-time forecasting systems.