Crop Yield Prediction

The **"Crop Yield Prediction"** project focuses on understanding agricultural production trends and forecasting future crop yields using historical data. Agriculture is the backbone of many economies, and improving crop productivity is essential for ensuring food security, optimizing land use, and supporting farmers.

In this project, I have analyzed agricultural data related to different crops, seasons, states, and years. By cleaning and preparing the data in **Excel**, querying it using **SQL**, visualizing patterns with **Power BI**, and building predictive models using **Python and Machine Learning**, we aim to extract valuable insights and predict crop yields for upcoming years (2025 to 2030).

Objectives of the Project:

- Clean and prepare raw agricultural datasets for analysis.
- Perform exploratory data analysis (EDA) to find key yield trends.
- Use **SQL** to answer important business and policy-related questions.
- Build dynamic and interactive dashboards in **Power BI** for visualization.
- Develop and train a **machine learning model** using Python to predict future yields.
- Evaluate the model's performance and forecast yield data for the years **2025 to 2030**.

Technologies and Tools Used:

- **Excel**: Initial cleaning, computation of yield, and static charts
- **SQL**: Querying, grouping, and filtering large volumes of crop data
- **Power BI**: Visual storytelling and interactive dashboards
- **Python** : EDA, modeling, forecasting using XGBoost
- **Machine Learning Algorithms**: Regression models for prediction

---

1. Excel : Data Cleaning & Preparation

Key Tasks Performed:

- **Initial Data Inspection**: Checked for null values, format inconsistencies, and structural errors.
- **Date Standardization**: Crop years were normalized (e.g., 2012–2013 as 2013).

- **Missing Data Handling**:
    - Production or Area missing → row removed or imputed with mean (depending on frequency).
- **Derived Column**: ○ Yield_Per_Hectare = Production / Area → essential for yield prediction.
- **Data Types Adjustment**:
    - State, Crop, and Season converted to categorical values.

Charts Created in Excel:

Chart Type	Description	Insight
Bar Chart	Crop vs. Yield	Identified high-yield crops (Wheat, Sugarcane)
Line Chart	Yearly Production Trend	Visualized rising/falling yield over years
Pie Chart	Seasonal Crop Distribution	Showed dominance of Kharif season
Stacked Column	State-wise Area vs. Production	Compared area-to-production ratio per state
Pivot Table	Aggregated yield per crop and year	Used for quick comparison and sorting

Summary:

- Visuals helped **identify trends**, **outliers**, and **high-performing states/crops**.
- Data exported for SQL analysis and ML pipeline.

SQL : Querying, Aggregation & All 10 Questions Solved

In the SQL part of the project, structured queries were written to extract meaningful insights from the crop dataset. The focus was on answering business and decision-making questions using SQL operations such as aggregation, filtering, grouping, and ranking.

SQL Questions:

1. **What is the average crop yield per state across all years?**
2. **Which crops have the highest average yield per hectare?**
3. **How has the average crop yield changed year by year?**
4. **What is the total production of a specific crop (e.g., Wheat) across all states?**
5. **What are the top 3 most produced crops in each state?**
6. **What is the total cultivated area grouped by season?**
7. **What is the year-wise trend of crop yield for a specific crop (e.g., Rice)?**
8. **Which state has the largest average cultivated area?**
9. **Which crop-state-year combination had the highest recorded yield?**
10. **Which crops contribute the most to total national production?**

---

3. Power BI – Visual Analytics & Dashboards

Pages Created:

1. **National Overview**
2. **Crop Insight**
3. **State-wise Dashboard**
4. **Seasonal View**

Visual Type	Description
Map	State-wise yield using color gradients
Bar Chart	Crop vs. average yield
Line Chart	Yearly yield trends
Pie Chart	Season-wise crop contribution
Table	Dynamic table with slicers for crop, state

Features:

- Interactive slicers: Filter by Crop, Year, State
- Drill-down: Click on a state to view only its data
- Conditional formatting: Yield thresholds color-coded Tooltips: Hovering on visuals shows detailed

---

4. Python : Analysis, Feature Engineering & Modeling

Data Processing :

- **Encoding**: LabelEncoded State, Season, and Crop.
- **Feature Generation**:
  - Yield_Per_Hectare as target.
  - Added lag features and moving averages (for smoothing).
- **Splitting**: Train-test (80-20) split.
- **Missing Data**: Imputed with mean or grouped averages.

EDA Visuals:

- **Heatmap**: Feature correlation
- **Boxplot**: Yield variation by crop
- **Line Plot**: Trend of yield over years

---

5. Machine Learning : Prediction & Evaluation

Model Used: XGBoost Regressor Evaluation:

- **R^2 Score**: ~0.87
- **MAE**: ~0.34
- **RMSE**: ~0.56

Forecasting (2025–2030):

- Created a synthetic dataset (states, crops, seasons) with average inputs for 2025–2030.
- Predicted yield for each year.
- Output visualized in: ○ Line plot (year-wise yield trend) ○ Heatmap (state vs. year)

Summary:

Step	Tool	Outcome
Data Prep	Excel	Cleaned and visualized data
SQL	SQL Server	Aggregated and queried key insights
Dashboard	Power BI	Visual analytics for stakeholders
Modeling	Python	Built and evaluated prediction model
Forecast	ML	Predicted yields for 2025–2030