Crop Yield Prediction using Machine Learning

Project Pitch Document

Prepared by: Veysel Berk

Date: October 13–25, 2025

Step 1: Business Problem

The Global Context

The global population has increased by over **25%** in the last two decades, but farmland has not kept up. According to the World Bank, agricultural land as a share of total land has declined from **37.2% in 2000 to 36.9% in 2022.** That might seem small, but at a global scale it reflects urban growth, land degradation, and climate pressure. Expanding farmland is no longer a realistic solution. It's costly, limited, and often harmful to the environment. That leaves one practical path forward: **increase yield from the land we already have**. This can be done through better crop selection, smarter input use, improved growing techniques, and data-driven decision-making. This project aims to support that goal by helping stakeholders understand the key drivers of crop yield.

1. Business Challenge

Crop yield is critical to food supply, economics, and planning. But many regions still rely on historical averages or experience to estimate yield. This project aims to:

- Predict expected yield using environmental and input features (rainfall, pesticide, temperature, etc.)

- Estimate production across countries and crops

- Identify what factors drive yield the most

2. Stakeholders

This project supports a range of stakeholders in the agriculture space:

- Growers looking to improve yield and income

- Agricultural advisors helping growers with input decisions

- Governments Agriculture Departments for planning food security for local production

- International organizations (like FAO, CGIAR) monitoring food security

- Ag-tech companies building decision tools and planning to enter new markets

- Researchers and NGOs studying climate and sustainability

3. Why It Matters

The global population is growing, but available farmland is limited. We need to produce more on the same land. Yield prediction helps:

- Growers make better choices

- Agencies allocate resources more efficiently

- Organizations track climate and policy impact on agriculture

4. Why Machine Learning

Traditional models miss complex patterns—like how rainfall interacts with crop type or pesticide levels. ML models can:

- Capture non-linear relationships

- Generalize across years, regions, and crops

- Provide scalable insights

5. Dataset Overview

- Source: [Crop Yield Prediction Dataset – Kaggle (Omdena project)]

- File: yield_df.csv

- Rows: 28,000+

- Key features: Area (country), Item (crop), Year, Yield (hg/ha), Rainfall (mm/year), Pesticide use (tonnes), Average temperature (°C)

6. Success Metrics

- Technical:

  - Low MAE and RMSE

  - R^2 above baseline

- Business:

  - Predictions within ±10–15% of actual yield

  - Clear identification of top yield drivers

Step 2: Problem Solving Process

Project Workflow

```
[ Business Understanding ]
            ↓
[ Data Understanding ]
            ↓
[ Data Preparation & Feature Engineering ]
            ↓
[ Modeling & Evaluation ]
            ↓
[ Interpretation & Recommendations ]
            ↓
[ Final Reporting & Submission ]
```

1. Data Understanding

- Load and inspect dataset

- Check for missing values and outliers

- Confirm units and ranges

- Explore feature distributions and trends

2. Data Preparation & Feature Engineering

- Drop unused columns

- Encode categorical features (Area, Item)

- Apply log transformation to skewed numeric features

- Create new features if needed (e.g., interactions)

- Build a reusable pipeline using scikit-learn

3. Modeling Strategy

- Baseline: Linear Regression

- Advanced Models: Random Forest, XGBoost or Gradient Boosting

- Use 5-fold cross-validation

- Tune models with GridSearchCV or RandomizedSearchCV

- Evaluate with MAE, RMSE, and R^2

- Use feature importance for interpretation

4. Results & Communication

- Visualize top yield drivers (by crop or region)

- Use clear plots: boxplots, bar charts, correlation heatmaps

- Translate findings into real-world language

  e.g., “In areas with high rainfall and mild temps, cassava shows higher yield”

- Connect results to business decisions

Step 3: Timeline and Scope

Project Duration: Oct 13–25, 2025

Phase 1 – Setup & Framing (1–2 days)

- Review data

- Finalize problem statement

- Set up GitHub repo

Phase 2 – Business Framing (1 day)

- Refine stakeholder framing

- Start on pitch document

Phase 3 – EDA (2–3 days)

- Profile features

- Explore trends and distributions

- Identify patterns and outliers

- Submit pitch document

Phase 4 – Data Prep (1–2 days)

- Clean and transform data

- Engineer features

- Build pipeline

Phase 5 – Modeling (2–3 days)

- Train baseline and advanced models

- Tune and cross-validate

Phase 6 – Evaluation (2–3 days)

- Final testing

- Visualize results

- Interpret model outputs

Phase 7 – Documentation (1-2 days)

- Finalize notebook

- Write README

- Save visuals

Phase 8 – Presentation & Submission (1–2 days)

- Create slides

- Record demo

- Final review and submission by Oct 25, 2025