Problem Understanding

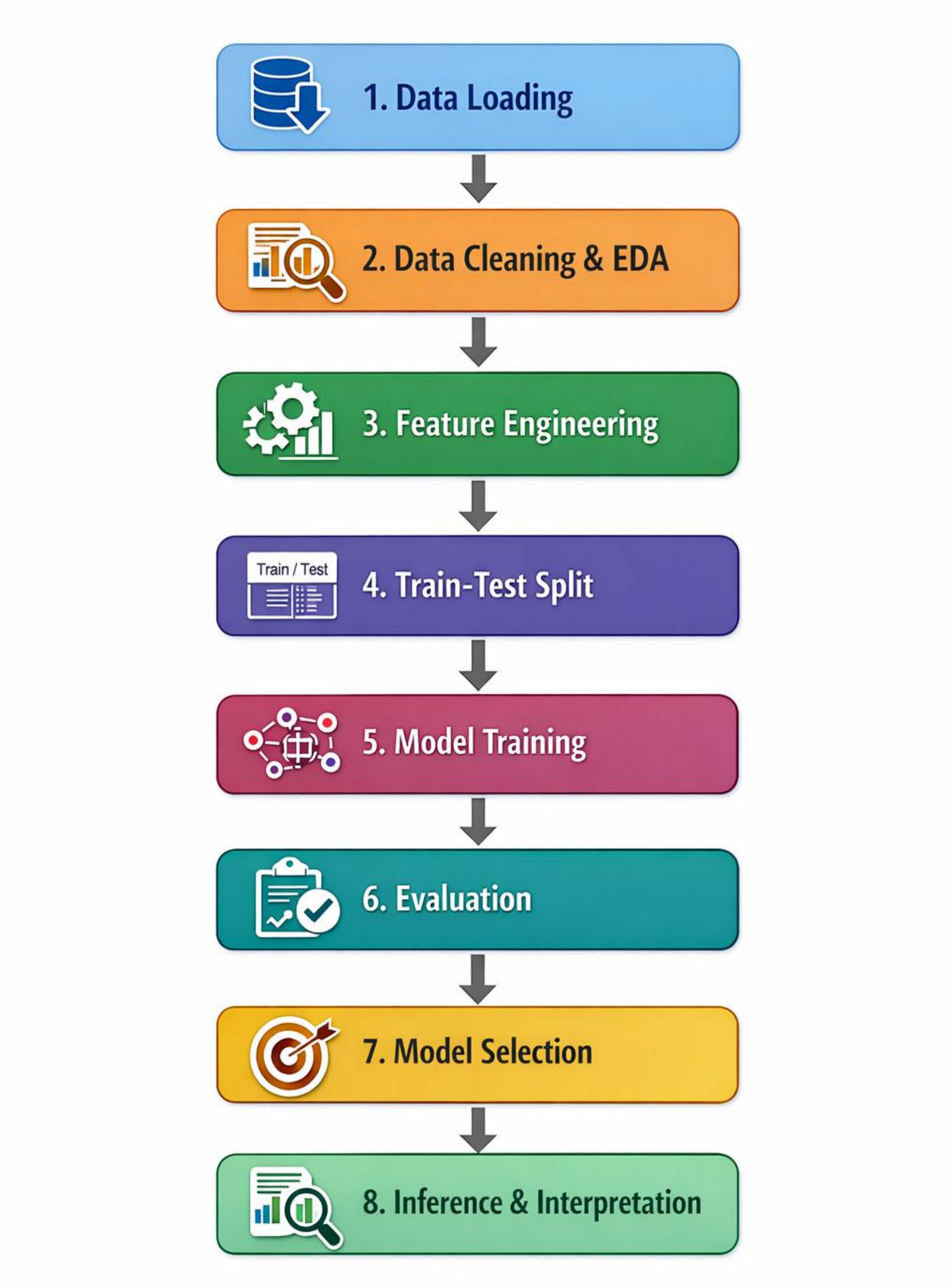
Predicting outcomes from real-world data is a key responsibility of an AI/ML engineer, as it enables data-driven decision making across industries. In the agricultural domain, crop yield prediction is a critical problem because it directly impacts food security, economic planning, and resource management. Crop yield depends on multiple environmental and agricultural factors, making it a suitable and realistic machine learning problem.

In this project, the selected problem is crop yield prediction, using a publicly available real-world dataset containing historical agricultural data. The dataset includes information such as country (Area), crop type (Item), year, average annual rainfall, pesticide usage, and average temperature. The target variable is crop yield measured in hectograms per hectare (hg/ha). These features represent real environmental and farming conditions that influence agricultural productivity.

The objective of the project is to build a machine learning model that can predict future crop yield based on historical patterns present in the data. By learning the relationship between environmental factors and crop yield, the model aims to provide accurate and reliable predictions for unseen data. Such predictions can help farmers, policymakers, and agricultural organizations make informed decisions regarding crop planning, irrigation, and resource allocation.

Rather than focusing only on achieving high accuracy, this project emphasizes proper data handling, model selection, and result interpretation. Multiple machine learning models are trained and evaluated to compare their performance and generalization ability. The final model is selected based on evaluation metrics and consistency across validation, and its predictions are interpreted using feature importance and error analysis to ensure transparency and reliability.

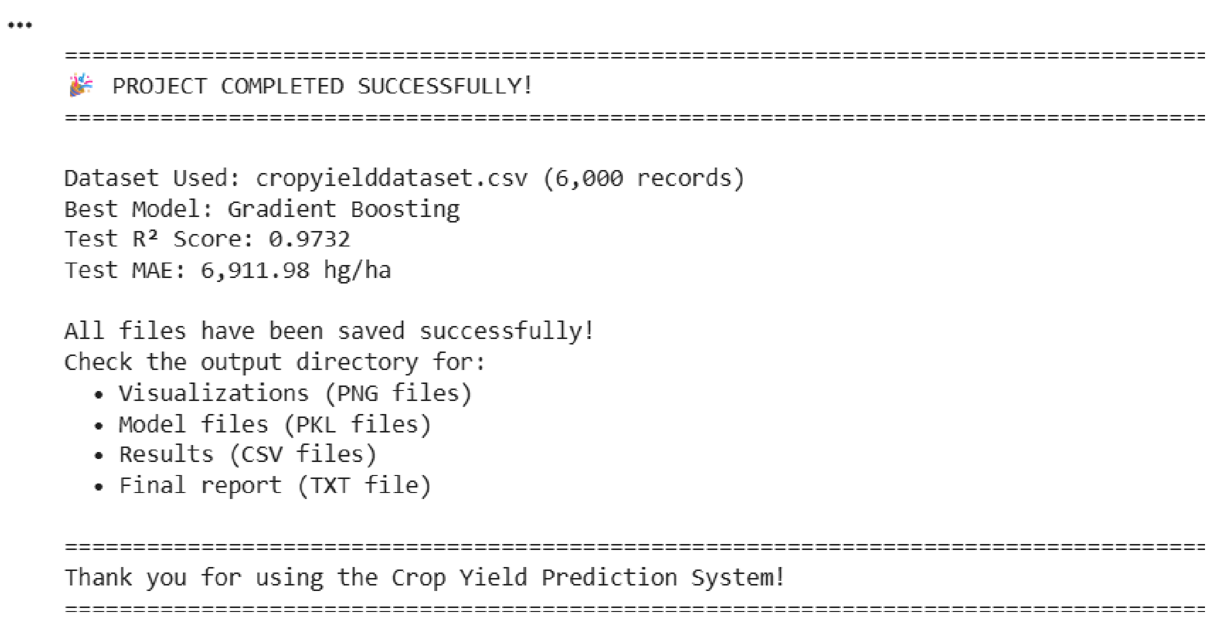
Model pipeline



Model Pipeline Description

1. The dataset was loaded from a CSV file and inspected for structure, missing values, duplicates, and data types.
2. Exploratory Data Analysis (EDA) was performed to understand crop yield distribution, trends across years, and relationships with rainfall and temperature.
3. Categorical features such as country (Area) and crop type (Item) were encoded using label encoding, while numerical features were scaled using StandardScaler.
4. The dataset was split into training (80%) and testing (20%) sets to evaluate generalization performance.
5. Multiple regression models including Linear Regression, Decision Tree, Random Forest, and Gradient Boosting were trained.
6. Models were evaluated using MAE, RMSE, and R^2 score on both training and test sets, along with 5-fold cross-validation.
7. The best model was selected based on test R^2 score and cross-validation consistency.
8. Feature importance and residual analysis were performed to interpret the model's predictions.

Results & metrics



```
================================================================================
🎉 PROJECT COMPLETED SUCCESSFULLY!
================================================================================

Dataset Used: cropyielddataset.csv (6,000 records)
Best Model: Gradient Boosting
Test R² Score: 0.9732
Test MAE: 6,911.98 hg/ha

All files have been saved successfully!
Check the output directory for:
  • Visualizations (PNG files)
  • Model files (PKL files)
  • Results (CSV files)
  • Final report (TXT file)

================================================================================
Thank you for using the Crop Yield Prediction System!
================================================================================
```