Climate Risk Modeling for Agriculture: Predicting Crop Yield

Your Name

August 2025

Abstract

This report presents a machine learning approach to predict crop yield (tons per hectare) using the `crop_yield.csv` dataset, addressing climate risk in agriculture. The project involves data loading, preprocessing, exploratory data analysis (EDA), feature engineering, model training, and evaluation. Key findings include the critical influence of rainfall and temperature on yield, with the best model achieving a test RMSE of `Your Test RMSE` and R^2 of `Your Test R2`. Recommendations for stakeholders include irrigation management and crop selection to mitigate climate risks.

1 Introduction

The objective of this project is to develop a predictive model for crop yield based on environmental and agricultural factors, aiding farmers and policymakers in managing climate risks. The dataset, `crop_yield.csv`, contains features such as `Rainfall_mm`, `Temperature_Celsius`, `Region`, and `Yield_tons_per_hectare`. The project follows a structured pipeline: data loading, preprocessing, EDA, feature engineering, model training, and evaluation.

2 Methodology

2.1 Data Loading and Preprocessing

The dataset was loaded using Python's `pandas` library. Non-numeric values (e.g., 'East' in numerical columns) were converted to NaN and imputed with column means. Missing categorical values were filled with the mode. Outliers beyond three standard deviations were removed. Numerical features were scaled using `StandardScaler`, and categorical features (`Region`, `Soil_Type`, `Crop`, `Weather_Condition`) were one-hot encoded using `OneHotEncoder`. The data was split into 70% training, 15% validation, and 15% test sets.

2.2 Exploratory Data Analysis (EDA)

EDA was conducted using `seaborn` and `matplotlib`. Histograms showed distributions of numerical features (`Rainfall_mm`, `Temperature_Celsius`, `Days_to_Harvest`). A correlation heatmap revealed relationships, e.g., `Rainfall_mm` strongly correlated with yield. Bar plots displayed average yield by categorical features, indicating regional and crop-specific trends.

2.3 Feature Engineering

New features were created to capture complex relationships:

- `Rainfall_Temperature_Interaction`: Product of `Rainfall_mm` and `Temperature_C` as their combined effect impacts yield.
- `Rainfall_mm_squared`, `Temperature_Celsius_squared`: Polynomial features to model non-linear effects.
- `High_Rainfall`: Binary indicator (1 if `Rainfall_mm` > median, 0 otherwise).

Feature importance was computed using a `RandomForestRegressor`, and features with importance above 0.01 were selected to reduce model complexity.

2.4 Model Training

Three regression models were trained using `scikit-learn` and `xgboost`:

- `LinearRegression`: Baseline model assuming linear relationships.
- `RandomForestRegressor`: Captures non-linear patterns, tuned with `GridSearchCV` (parameters: `n_estimators`, `max_depth`, `min_samples_split`).
- `XGBRegressor`: Gradient boosting for complex patterns.

Models were evaluated on the validation set using RMSE and R^2. The best model, `Your Best Model Name`, was selected based on the lowest RMSE.

2.5 Model Evaluation

The best model was evaluated on the test set, achieving an RMSE of `Your Test RMSE` and R^2 of `Your Test R2`. A histogram of prediction errors showed a mean error of `Your Mean Error`, indicating no significant bias. Feature importance (or coefficients) highlighted `Rainfall_mm`, `Temperature_Celsius`, and their interaction as key predictors.

3 Results

The best model, `Your Best Model Name`, achieved a test RMSE of `Your Test RMSE` tons/hectare and an R^2 of `Your Test R2`, indicating strong predictive power. Key findings:

- Rainfall and temperature are critical drivers of yield, with optimal ranges maximizing output.
- Regional and crop-specific variations suggest tailored agricultural strategies.
- The model is unbiased (mean error near zero), though some outliers indicate complex cases.

Test predictions were saved to `test_predictions.csv`.

4 Discussion

The model provides reliable yield predictions, enabling stakeholders to mitigate climate risks. Recommendations include:

- **Irrigation Systems**: Stabilize rainfall variability, especially in low-rainfall regions.
- **Crop Selection**: Choose varieties suited to local temperature and soil conditions.
- **Weather Monitoring**: Adjust planting schedules based on forecasts.
- **Soil Management**: Enhance soil quality to improve resilience.

5 Conclusion

This project successfully developed a machine learning model to predict crop yield, leveraging Python, `pandas`, `scikit-learn`, `xgboost`, `seaborn`, and `matplotlib`. The model's performance and insights provide actionable strategies for optimizing agricultural yield under climate variability. Future work could explore additional features or advanced models like neural networks.

6 Tools and Technologies

- **Python**: Programming language for data processing and modeling.
- **pandas**: Data loading, cleaning, and manipulation.
- **scikit-learn**: Preprocessing, model training, and evaluation.

- **xgboost**: Gradient boosting model.
- **seaborn, matplotlib**: Data visualization.
- **Jupyter Notebook**: Interactive development environment.