PREDICTING CROP YIELDS USING MACHINE LEARNING

Gregor Pfister |MIS | Wintersemester 2024/25

MOTIVATION AND OBJECTIVE

Agriculture productivity is essential for food security.

Understanding how environmental factors influence crop yield.

Objective: Explore if crop yield can be accurately predicted using machine learning techniques.

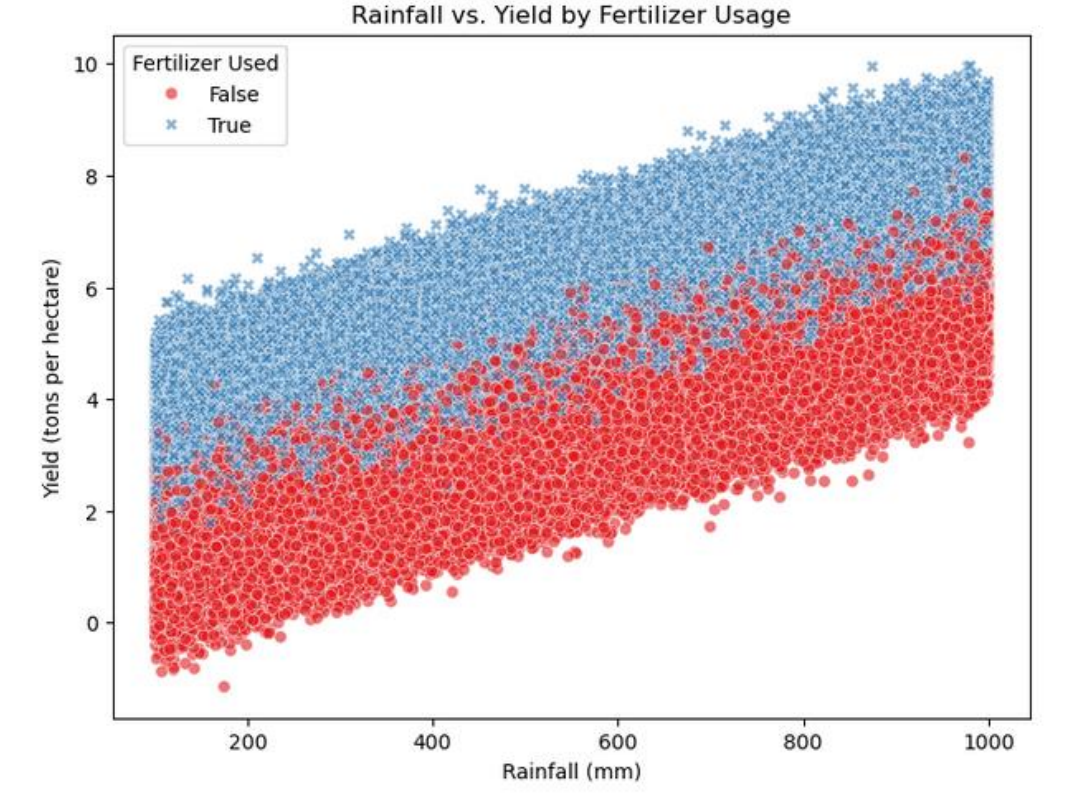
RESEARCH QUESTION

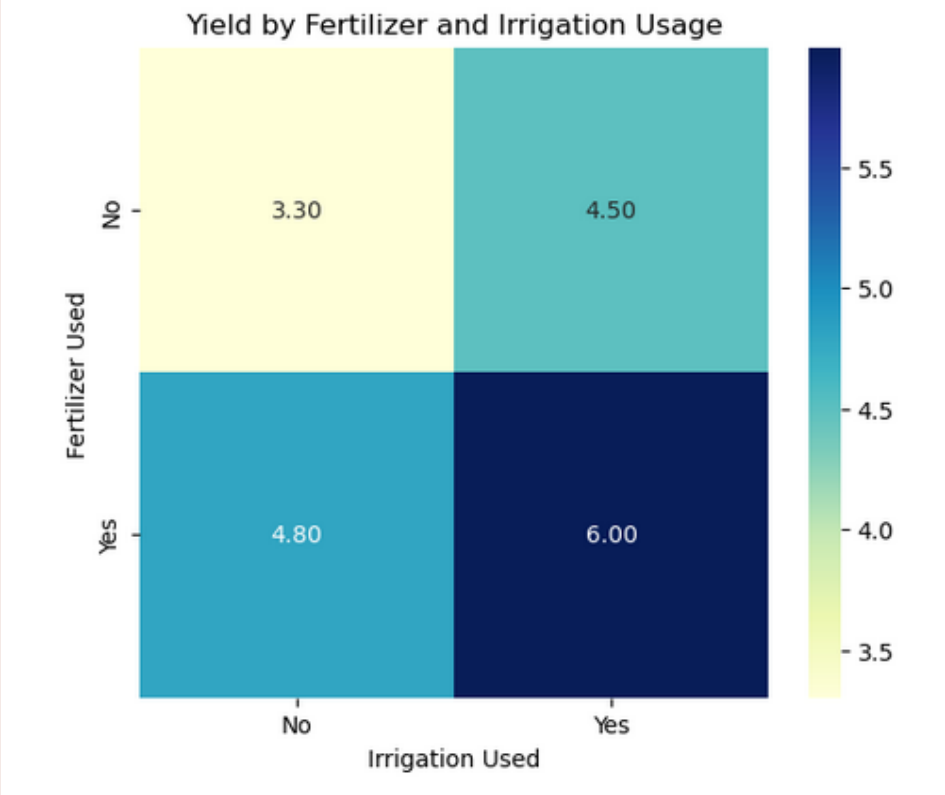
"Can crop yields be predicted with environmental factors like rainfall, soil type or fertilization"

DATASET DESCRIPTION

- Agriculture Crop Yield Dataset
- **Source**: https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield?resource=download
- **Size**: 1,000,000 rows × 10 columns
- **Key Features**:
  - Rainfall (mm), Fertilizer Usage, Irrigation, Temperature (°C)
  - Yield (tons per hectare) as the target variable
- **Goal**: Predict yield based on environmental and management factors.

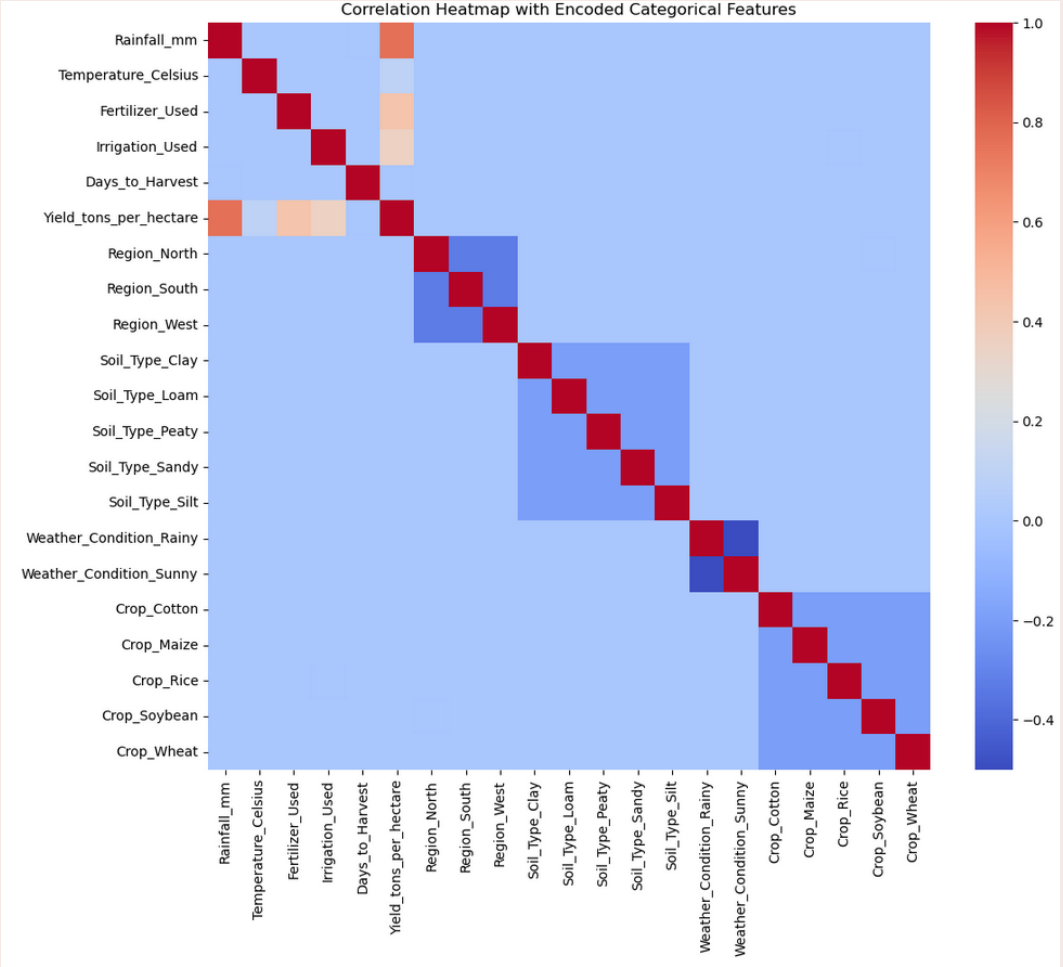
INITIAL VISUALIZATIONS





KEY CORRELATIONS

- Rainfall, Fertilizer Usage, and Irrigation have the highest correlation with crop yield.
- Rainfall is the dominant predictor (correlation: **0.76**).



AUTOML WITH AUTOGLUON

- Preprocessing: Feature selection, train-test split.
- Model Training: AutoGluon Tabular predictor experiments.
- Metrics: RMSE, MAE, R^2, Accuracy (%).

AUTOML WITH AUTOGLUON

- First Experiment: Training on the 3 identified key features

- **Root Mean Squared Error (RMSE)**: 0.521 tons/ha
- **Mean Squared Error (MSE)**: 0.271 $tons^2$/ha
- **Mean Absolute Error (MAE)**: 0.416 tons/ha
- **Median Absolute Error**: 0.352 tons/ha
- **R^2 Score**: 0.906
- **Pearson Correlation Coefficient**: 0.952

Feature	Importance	StdDev	p-value	p99 High	p99 Low
Rainfall (mm)	1.39	0.005	2.06e-11 (significant)	1.40	1.38
Fertilizer Used	0.66	0.004	1.48e-10 (significant)	0.67	0.65
Irrigation Used	0.47	0.007	4.77e-09 (significant)	0.48	0.46

AUTOML WITH AUTOGLUON

- Second Experiment: Full Dataset without feature limitation

Metric	Key Features Model	Full Dataset Model
Root Mean Squared Error	0.521	0.501
Mean Squared Error	0.271	0.251
Mean Absolute Error	0.416	0.400
Median Absolute Error	0.352	0.337
R^2 Score	0.906	0.913
Pearson Correlation (r)	0.952	0.955

Feature	Importance	StdDev	p-value
Rainfall (mm)	1.406	0.005	2.90e-11
Fertilizer Used	0.671	0.004	2.37e-10
Irrigation Used	0.481	0.008	8.08e-09
Temperature (°C)	0.037	0.002	1.35e-06
Other Features	Negligible	-	-

AUTOML WITH AUTOGLUON

- Third Experiment: Training on the 4 key features identified in the last experiment

Metric	Key Features Model	Selected Features Model	Full Dataset Model
Root Mean Squared Error	0.521	0.501	0.501
Mean Squared Error	0.271	0.251	0.251
Mean Absolute Error	0.416	0.400	0.400
Median Absolute Error	0.352	0.337	0.337
R^2 Score	0.906	0.913	0.913
Pearson Correlation (r)	0.952	0.955	0.955

AUTOML WITH AUTOGLUON

- Fourth Experiment: Key Features + custom hyperparameters

Metric	Custom HPO Model	Selected Features Model	Full Dataset Model	Key Features Model
Root Mean Squared Error	0.501	0.501	0.501	0.521
Mean Squared Error	0.251	0.251	0.251	0.271
Mean Absolute Error	0.400	0.400	0.400	0.416
Median Absolute Error	0.338	0.337	0.337	0.352
R^2 Score	0.913	0.913	0.913	0.906
Pearson Correlation (r)	0.955	0.955	0.955	0.952

BEST MODEL: SELECTED FEATURES

Lowest RMSE (0.501) and highest Accuracy (91.41%).

Features: Rainfall, Fertilizer Usage, Irrigation, Temperature.

Balance between simplicity and performance.

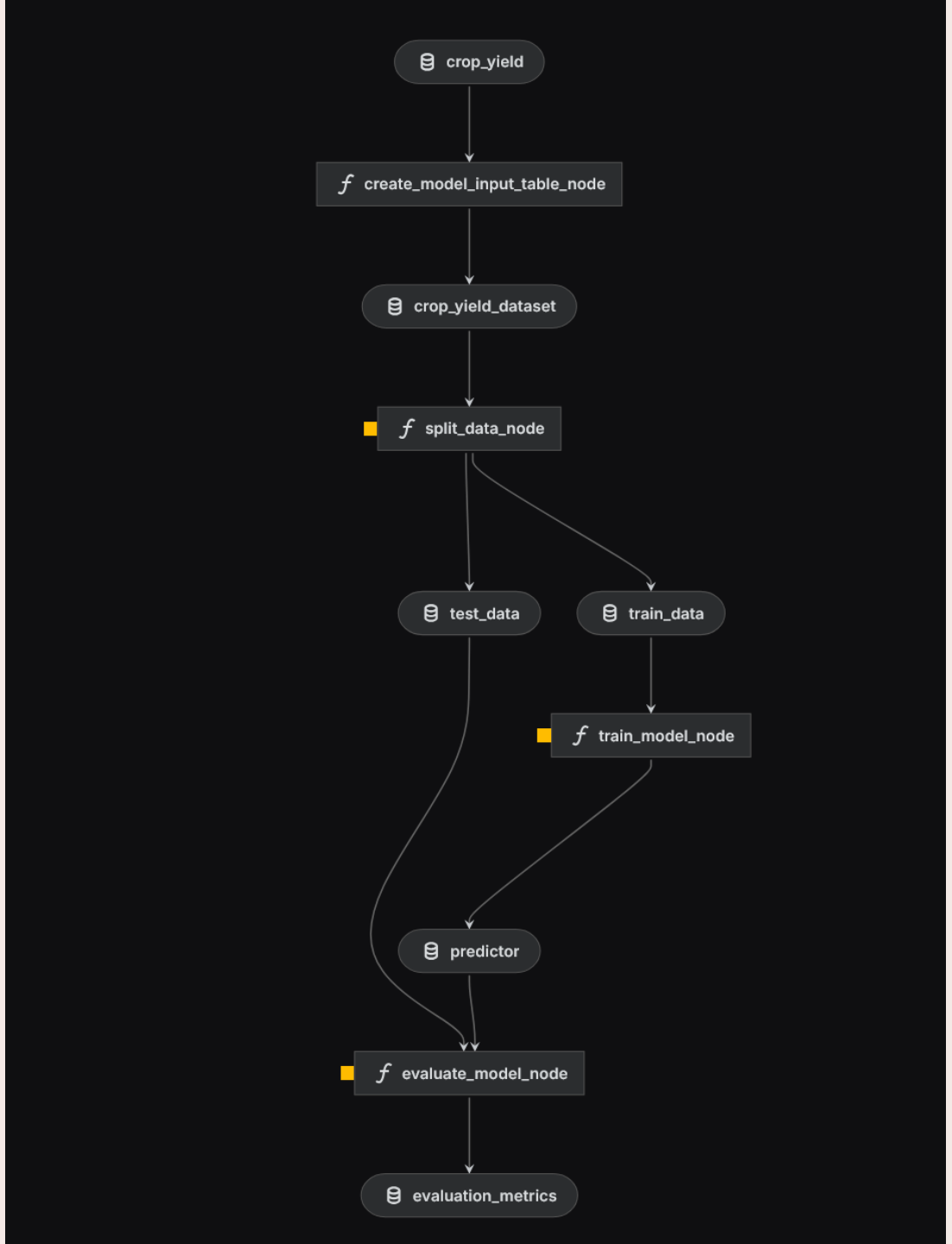
CHALLENGES AND INSIGHTS

Minimal performance gain from adding features or tuning.

Rainfall is overwhelmingly dominant; other features had minor effects.

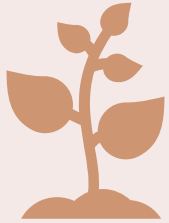
FUTURE WORK

- Kedro pipeline: WIP
- More experiments with custom hyperparameter tuning

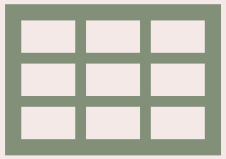


KEY TAKEAWAYS

Machine learning can accurately predict crop yield (R^2 = 0.91+).



Rainfall, Fertilizer, and Irrigation are critical features.



Simple models (Selected Features) perform as well as complex ones.

ANY QUESTIONS?