Data-Driven Agriculture

Cheyenne Edwards, Parneet Kaur, Sowmya Renukuntla

Table of contents

Overview

- **Definition:** a standard measurement of the amount of agricultural production harvested per unit of land area
  - **Unit of Measurement:** hg/ha (hectograms per hectare)
- **Importance/Motivation:**
  - Improving food security
  - Issue: Farmers face the challenge of identifying which crops will yield the highest returns in a given region.
    - Ultimate Goal - Recommendation System that:
      - Analyzes geographic environmental factors
      - Finds similar regions
      - Recommends high-yield crops
      - Predicts the expected yield using a supervised learning model

02

Sub-Questions/ Hypotheses

Central Question

What are the key features that best predict crop yield?

Sub-Question #1:

How does adding environmental factors like rainfall and temperature affect a model's ability in predicting crop yield?

- **Data Visualization/Exploration:**
  - Heatmap for Correlation Analysis
  - Scatter Plots and Polynomial Regression
  - 2D Heatmap
  - OLS (Ordinary Least Squares) Regression
    - Establish a baseline model
    - Quantifies how much rainfall and temperature affect crop yield through:
      - R-Squared (R^2)
      - p -Value
- **Models:**
  1. Baseline Model - Time trends and crop differences without environmental factors.
  2. Extended Model with Temperature & Rainfall
  3. Quadratic Model
  4. Interaction Terms Model

Sub-Question #2:

How does adding pesticide usage as a feature affect the model's ability in predicting crop yield?

- **Data Visualization**
  - Correlation Analysis
    - Pesticide usage has a low correlation with crops and yield
  - Scatterplot
    - Suggested that pesticide usage alone may be the sole factor influencing crop yield
- Models
  - Linear and polynomial regression
  - Random Forest
  - XGBoost Ensemble

Holistic Models

- **Local weak models (focused on specific factors) -> Holistic Models**
  - Environmental factors (pesticide, temperature, and rainfall) were weak predictors + did not achieve satisfying performance metrics
- **Transition to tree-based models** in order to:
  1. Increase R^2
     a. Goal: Get a model to explain a large proportion of the variance in the dependent variable (i.e. Yield) with minimal to no overfitting
  2. Test for feature importance
     a. Answers Central Question
- **Models**
  1. Decision Tree
  2. Random Forest
  3. XGBoost (Ensemble Learning Method)

Sub-Question #3:

Does training separate models for different clusters (regional models) lead to better predictive performance than a single global model?

- Don't want to overgeneralize important patterns
  - Could reduce predictive accuracy
- Compare to a global model < clustered variable model
  - Lead to higher predictive accuracy

Models:

1. K-Means Clustering (environmental conditions -> crop yields)
2. Linear Regression (change in dummy variable -> change in yield)
3. Random Forest (ranking of dummy variable -> regional differences affect on yield predictions)

03

Analysis

Dataset Content

Four datasets: Pesticides, Rainfall, Temperature, Yield

Year: the year the data was collected

Avg_Temp: the average temperature for a certain year

Latitude/Longitude: the country data was collected from

Average_Rainfall_Per_Year: average rainfall by meter

Value_Pesticide: amount of pesticides used (tonnes of active ingredients)

Crop_Type: the kind of crop (maize, rice, potatoes, etc.)

Crop_Yield: the yield amount of the crop

Datasets Preview

Yield Data:

	Domain Code	Domain	Area Code	Area	Element Code	Element	Item Code	\
0	QC	Crops	2	Afghanistan	5419	Yield	56	
1	QC	Crops	2	Afghanistan	5419	Yield	56	
2	QC	Crops	2	Afghanistan	5419	Yield	56	
3	QC	Crops	2	Afghanistan	5419	Yield	56	
4	QC	Crops	2	Afghanistan	5419	Yield	56	

	Item	Year Code	Year	Unit	Value
0	Maize	1961	1961	hg/ha	14000
1	Maize	1962	1962	hg/ha	14000
2	Maize	1963	1963	hg/ha	14260
3	Maize	1964	1964	hg/ha	14257
4	Maize	1965	1965	hg/ha	14400

Temperature Data:

	year	country	avg_temp
0	1849	Côte D'Ivoire	25.58
1	1850	Côte D'Ivoire	25.52
2	1851	Côte D'Ivoire	25.67
3	1852	Côte D'Ivoire	NaN
4	1853	Côte D'Ivoire	NaN

Rainfall Data:

	Area	Year	average_rain_fall_mm_per_year
...			
1	tonnes of active ingredients		121.0
2	tonnes of active ingredients		121.0
3	tonnes of active ingredients		121.0
4	tonnes of active ingredients		201.0

Pesticides Data:

	Domain	Area	Element	Item	Year	\
0	Pesticides Use	Albania	Use	Pesticides (total)	1990	
1	Pesticides Use	Albania	Use	Pesticides (total)	1991	
2	Pesticides Use	Albania	Use	Pesticides (total)	1992	
3	Pesticides Use	Albania	Use	Pesticides (total)	1993	
4	Pesticides Use	Albania	Use	Pesticides (total)	1994	

	Unit	Value
0	tonnes of active ingredients	121.0
1	tonnes of active ingredients	121.0
2	tonnes of active ingredients	121.0
3	tonnes of active ingredients	121.0
4	tonnes of active ingredients	201.0

Basic Preprocessing & Merging

- **Previewed the data first** even though there were four separate datasets
  - Understanding the structure for **yield_df, rain_df, temp_df, pest_df**
    - Final Dataset Shape: (28119, 19)
    - Checking for columns that have **single unique value**
      - Applying **Filter-Based Feature Selection** (if needed)
  - **Identifying Missing Values** (isnull() was not enough)
  - **Identifying Duplicate Rows**
    - duplicated() - Identified duplicate rows across the 'Year' column
- **Merging Datasets -** Left Merge sequentially
  - Yield Data (yield_df) → merged with Rainfall Data (rain_df)
  - Merged dataset → merged with Pesticide Data (pest_df)
  - The updated dataset → merged with Temperature Data (temp_df)

1. Identifying Missing Values

- **Datasets Containing Missing Values:**
  - temp_df: avg_temp -> 2547 missing
  - rain_df: Average_Rainfall_Per_Year -> 774 missing
- **Goal: Minimize bias**
  - **Identified Missing Data Patterns**
    - Missing temperature values were mainly occurring in specific regions -> MAR
  - **Handled Missing Data Based on Proportions**
    - <= 30% Missing: Used interpolation by country
    - 30%-70% Missing: Applied KNN Imputation
    - >70% Missing: Dropped
  - **Special Case: rain_df**
    - Missing Not at Random (MNAR)

```
American Samoa                 31
Aruba                          31
Turks and Caicos Islands       31
Tonga                          31
St. Martin (French part)       31
Sint Maarten (Dutch part)      31
San Marino                     31
Northern Mariana Islands       31
New Caledonia                  31
Macao SAR, China               31
Kosovo                         31
Isle of Man                    31
Hong Kong SAR, China           31
Guam                           31
Greenland                      31
Gibraltar                      31
French Polynesia               31
Faroe Islands                  31
Curacao                        31
Channel Islands                31
Cayman Islands                 31
British Virgin Islands         31
Bermuda                        31
Virgin Islands (U.S.)          31
Monaco                         30
Name: Area, dtype: int64
```

```
      Area Average_Rainfall_Per_Year
4061  Monaco                      ..
```

2. Identifying Duplicate Rows

- Only had duplicate rows in temp_df
- **Method:**
  - Identified duplicate rows with different temperature values for the same year and area.
  - Instead of dropping them, we computed the average temperature.
    - Different temperature values indicate multiple recordings
  - One representative value per year for each area

	Year	Area	avg_temp
0	1849	Côte D'Ivoire	25.580
1	1850	Côte D'Ivoire	25.520
2	1851	Côte D'Ivoire	25.670
3	1852	Côte D'Ivoire	25.792
4	1853	Côte D'Ivoire	25.914
...	...	...	...
71306	2009	Mexico	21.760
71307	2010	Mexico	20.900
71308	2011	Mexico	21.550
71309	2012	Mexico	21.520
71310	2013	Mexico	22.190

71311 rows × 3 columns

	Year	Area	avg_temp
0	1849	Côte D'Ivoire	25.58
70465	1849	Côte D'Ivoire	25.00

	Area	Year	avg_temp
6476	Côte D'Ivoire	1849	25.29

3. Encoding

- Performed longitude-latitude encoding since we have countries in dataset
- Can help model understand geographical location of each country and its relationship with crop yield

	Area	Item_x	Year	Unit_x	Value_Yield	Average_Rainfall_Per_Year	Value_Pesticide	avg_temp	Category
0	Afghanistan	Maize	1990	hg/ha	17582	1130.0	1597.0	24.55	Other
1	Afghanistan	Maize	1991	hg/ha	16800	1130.0	1597.0	24.55	Other
2	Afghanistan	Maize	1992	hg/ha	15000	1130.0	1597.0	24.55	Other
3	Afghanistan	Maize	1993	hg/ha	16786	1130.0	1597.0	24.55	Other
4	Afghanistan	Maize	1994	hg/ha	16667	1130.0	1597.0	24.55	Other

	Area	Crop Name	Year	Value_Yield	Average_Rainfall_Per_Year	Value_Pesticide	avg_temp	Category	latitude	longitude
0	Afghanistan	Maize	1990	17582	1130.0	1597.0	24.55	Other	33.93911	67.709953
1	Afghanistan	Maize	1991	16800	1130.0	1597.0	24.55	Other	33.93911	67.709953
2	Afghanistan	Maize	1992	15000	1130.0	1597.0	24.55	Other	33.93911	67.709953
3	Afghanistan	Maize	1993	16786	1130.0	1597.0	24.55	Other	33.93911	67.709953
4	Afghanistan	Maize	1994	16667	1130.0	1597.0	24.55	Other	33.93911	67.709953

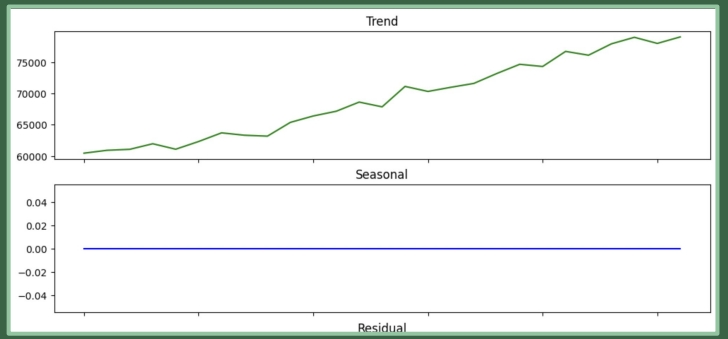
- Performed OneHotEncoding on Category and Crop columns to create an artificial ordinal relationship

#	Column
0	Category_SIDS
1	Category_SIDS_LDC
2	Crop_Maize
3	Crop_Plantains_Others
4	Crop_Potatoes
5	Crop_Rice_Paddy
6	Crop_Sorghum
7	Crop_Soybeans
8	Crop_Sweet_Potatoes
9	Crop_Wheat
10	Crop_Yams
11	Area
12	Year
13	Value_Yield
14	Average_Rainfall_Per_Year
15	Value_Pesticide
16	Avg_Temp
17	Latitude
18	Longitude

04

Findings

Sub-Question #1



Data Exploration

- Time-Dependent Patterns -> Time Series Forecasting Models (e.g. SARIMA)
  - Time is an important predictor - Usage in Linear Models
  - BUT, seasonal component is flat (zero)
- **OLS Regression -** Primarily used to check for quadratic trend between temperature, rainfall, and yield
  - Included $Temperature^2$ and $Rainfall^2$ to test
    - Previous Observations from Polynomial Curve: U-Shaped Pattern

Variable	Coefficient	t-Statistic	p-value	Significance
Avg_Temp	-5.419e+05	-4.965	0.000	Significant
Avg_Temp_Square	1.04e+04	4.937	0.000	Significant
Avg_Rainfall_Per_Year	2.6085	0.364	0.716	Not Significant
Rainfall_Squared	0.0014	0.709	0.479	Not Significant

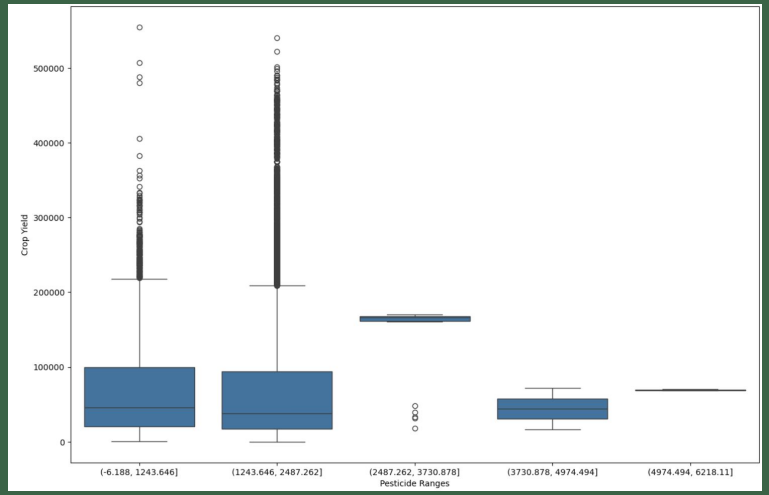
Modeling

- **Time-aware train/test sets**
  - Training = 1990-2010 (~77.4% of data)
  - Testing = 2011–2016 (~22.6%)

Models / Formula	R^2	Train RMSE
$Yield = \beta0 + \beta1 \times Year + \beta2 \times Crop\ Type + \epsilon$	Train: 0.4887 Test: 0.4967	52458.95
$Yield = \beta0 + \beta1 \times Year + \beta2 \times Crop\ Type + \beta3 \times Rainfall + \beta4 \times Temperature + \epsilon$	Train: 0.4938 Test: 0.5036	52194.29
$Yield = \beta0 + \beta1 \times Year + \beta2 \times Crop\ Type + \beta3 \times Rainfall + \beta4 \times Temperature + \beta5 \times Temperature^2 + \epsilon$	Train: 0.4948 Test: 0.5036	52141.60
$Yield = \beta0 + \beta1 \times Year + \beta2 \times Crop\ Type + \beta3 \times Rainfall + \beta4 \times Temperature + \beta5 \times Temperature^2 + \beta6(Temperature \times Rainfall) + \epsilon$	Train: 0.5113 Test: 0.5238	51287.21

Sub-Question #2

- Created a time-series split for training and testing data
- Found that pesticide usage is not a key feature in predicting yield
- **Linear Regression**
  - R^2 = 0.004 with pesticide included as a feature
- **Polynomial Regression**
  - R^2 = 0.04 with pesticide included as a feature
  - Increase in both baseline and pesticide set indicates non-linear relationship
- Utilized more advanced models, but the simple ones reinforce the findings




```
Linear Regression MSE without Pesticide: 3380799850.5882
Linear Regression RMSE without Pesticide: 58144.6459
Linear Regression R-squared without Pesticide: 0.5085

Linear Regression MSE with Pesticide: 6848136926.6793
Linear Regression RMSE with Pesticide: 82753.4708
Linear Regression R-squared with Pesticide: 0.0045
```

```
Polynomial Regression MSE without Pesticides: 2972510848.1511
R2 without Pesticides: 0.5679

Polynomial Regression MSE with Pesticides: 6562432301.4386
R2 without Pesticides: 0.0460
```

Holistic Models - Performance

- Standard train-test split (Year ≤ 2010 for training, Year > 2010 for testing)
  - **TimeSeriesSplit** -> sequential time-based folds (e.g., train on 1990-1995, validate on 1996-2000, etc.).
- **Hyperparameter Tuning** - GridSearchCV & RandomizedSearchCV
- **XGBoost** - Separate Cross Validation
  - XGBoost optimize based on previous errors and may require additional fine-tuning
    - Tuned n_estimators by running multiple folds and stopping when RMSE stopped improving.

Model	Training R^2	Testing R^2	Training RMSE	Testing RMSE
Decision Tree	0.8965	0.8139	23,601.80	35,780.76
Random Forest	0.9412	0.8549	17,791.66	31,589.29
Random Forest (Log Transformation on Yield)	0.9429	0.8788	21,480.24	36,667.81
XGBoost	0.9084	0.8388	22,203.27	33,300.14

Holistic Models - Feature Importance

- Explored why we got these results

Is the model giving high importance to Potatoes because their yields vary a lot (have high variance), making them naturally useful for the tree to split on?

```
Crop_Maize Yield Variance: 1304633970.97
Crop_Plantains_Others Yield Variance: 5264517493.37
Crop_Potatoes Yield Variance: 9404786031.26
Crop_Rice_Paddy Yield Variance: 348717035.52
Crop_Sorghum Yield Variance: 345680817.53
Crop_Soybeans Yield Variance: 56348024.20
Crop_Sweet_Potatoes Yield Variance: 5066609962.25
Crop_Wheat Yield Variance: 345359539.10
Crop_Yams Yield Variance: 2082524695.10
```

Decision Tree

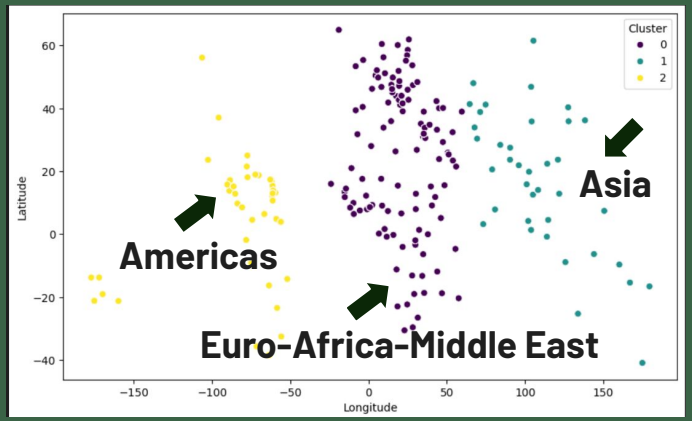
	Feature	Importance
4	Crop_Potatoes	0.360653
15	Latitude	0.219539
16	Longitude	0.203866
8	Crop_Sweet_Potatoes	0.048825
5	Crop_Rice_Paddy	0.042686
10	Crop_Yams	0.022474
3	Crop_Plantains_Others	0.021570
11	Year	0.017138
2	Crop_Maize	0.012988
6	Crop_Sorghum	0.011804
7	Crop_Soybeans	0.010012
9	Crop_Wheat	0.009636
13	Value_Pesticide	0.005774
12	Average_Rainfall_Per_Year	0.004804
17	Avg_Temp_Squared	0.004503
14	Avg_Temp	0.001884
1	Category_SIDS_LDC	0.001011
0	Category_SIDS	0.000833

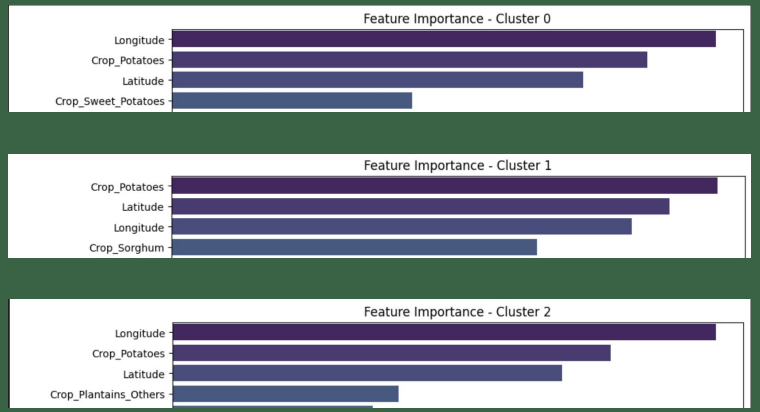
Random Forest

	Feature	Importance
0	Crop_Potatoes	0.183951
1	Latitude	0.178339
2	Longitude	0.142774
3	Crop_Sorghum	0.111435
4	Crop_Soybeans	0.103761
5	Crop_Maize	0.065074
6	Crop_Wheat	0.055145
7	Crop_Rice_Paddy	0.039987
8	Crop_Sweet_Potatoes	0.031439
9	Year	0.021195
10	Average_Rainfall_Per_Year	0.016391
11	Crop_Yams	0.014812
12	Crop_Plantains_Others	0.012300
13	Avg_Temp_Squared	0.006537
14	Avg_Temp	0.006363
15	Value_Pesticide	0.005869
16	Category_SIDS	0.003013
17	Category_SIDS_LDC	0.001615

Sub-Question #3

- Silhouette score recommended k=3
- PCA Feature Importance -> longitude and latitude
  - Primary drivers of variance





- Each region has distinct drivers (i.e crop, agronomic, etc.)
- Clustered model is much more interpretable than a global model

Recommendation System

- **Goal:**
  - Best crop to grow in a given geographical and environmental condition
  - Estimated yield for the recommended crop
  - Additional crop suggestions based on similar regions.
- **Content-Based Recommendation System**
  - Had to run K-Means again on copy of original dataframe
  - Assigned test sample to one of 3 precomputed clusters
    - Counted frequency of each crop in the cluster (Most Frequently Grown -> Recommended)
  - Used pre-trained Random Forest models for each cluster to predict new yield
  - Back-Up Crop(s) Recommended using similarity matrix
    - Cosine Similarity applied
    - Excluded the already recommended crop
    - Provided yield estimate

Recommendation System

```
test_sample = [37.7749, -122.4194, 26, 1200, 50]
```

- **Latitude/Longitude** - San Francisco, CA
- **Other test inputs** - randomly created

What did we expect? - Potatoes

Best Crops Per Region

	Cluster	Latitude	Longitude	Value_Yield	Crop_Maize	Crop_Plantains_Others	Crop_Potatoes
1903	0	50.503887	4.469936	540003	0.0	0.0	1.0
17719	1	-40.900557	174.885971	495751	0.0	0.0	1.0
2026	2	17.189877	-88.497650	554855	0.0	1.0	0.0

Results

```
Recommended Crop: Potatoes
Estimated Yield: 111683.41 hg/ha
Additional Crop Suggestions and Estimated Yields: {'Wheat': 48022.078783837576}
```

Moving Forward


Thank You?
Any questions?

Appendix

Exhibit A: Outlier Detection