Precision Agriculture: Addressing the Global Food Crisis

Medha Boddu, Brendan Puglisi, Ronica Peraka

Question

Given a set of agricultural conditions, how can we create a ML model that predicts crop yield?

Background

- **Problem**
  - According to the World Food Programme (WFP), more than 48 million people are facing emergency levels of hunger, with the threat of acute malnutrition, starvation, and death.
    - The United Nations reported the need for a 60% increase in food production by *2050* if we hope to serve our growing population of 9.7 billion
  - Global warming has begun to affect local climates changing rainfall patterns as well as soil conditions rendering farmers once superior local knowledge increasingly ineffective
- **Solution**
  - **Precision Agriculture:** Integrating Data Science into Farming
    - Aerial imagery and AI image identification to monitor and manage fields
    - Crop monitors to efficiently use input resources
    - Soil monitors identifying best crop recommendations

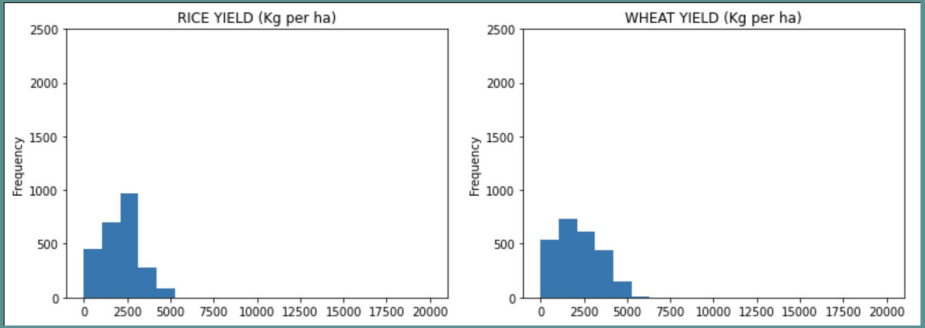
Reid, Kathryn. "10 World Hunger Facts You Need to Know." *World Vision*, 30 Aug. 2022, www.worldvision.org/hunger-news-stories/world-hunger-facts.

"Feeding the World Sustainably." *United Nations*, United Nations, www.un.org/en/chronicle/article/feeding-world-sustainably#:~:text=According%20to%20estimates%20compiled%20by,world%20population%20of%209.3%20billion.

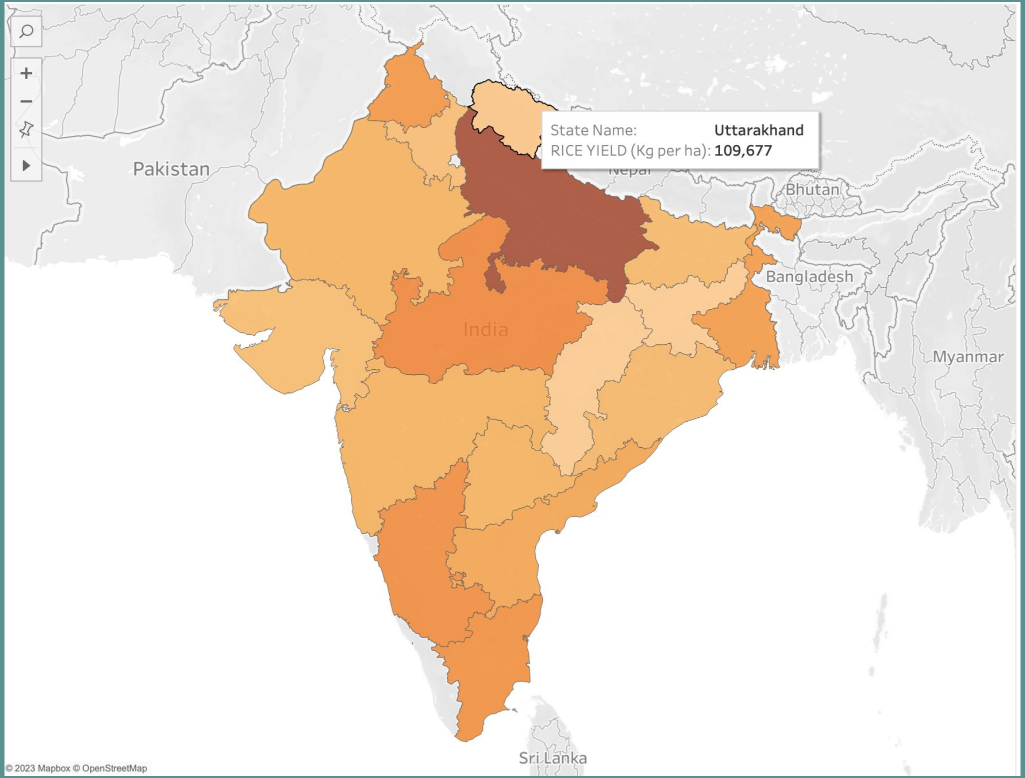
Exploratory Data Analysis

Data Cleaning

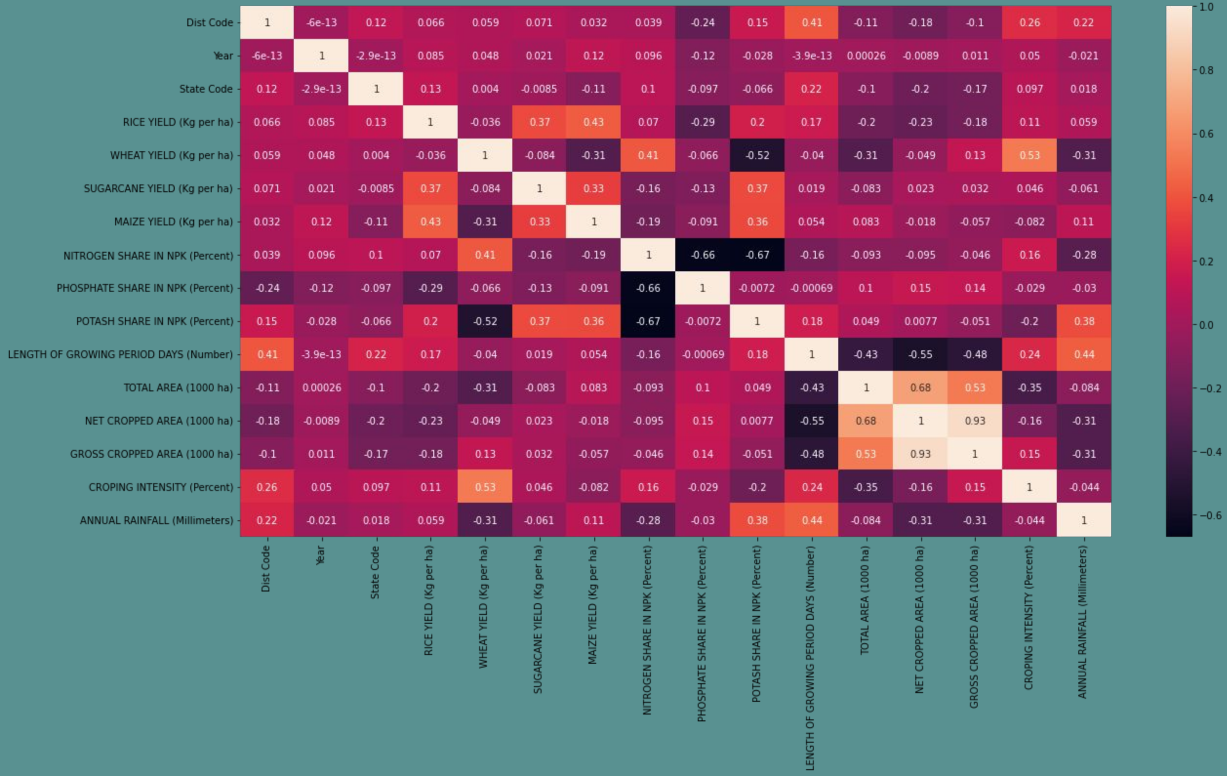
- Merged 5 datasets using district code
- All datasets contained environmental conditions that impact crop yield
- Selected Rice Yield and Wheat Yield (kg per ha) due to distributions
- Converted into a long format to create a column for crop type
- Replaced NAs with the mean for Annual Rainfall and Cropping Intensity
- Removed predictors with high correlation



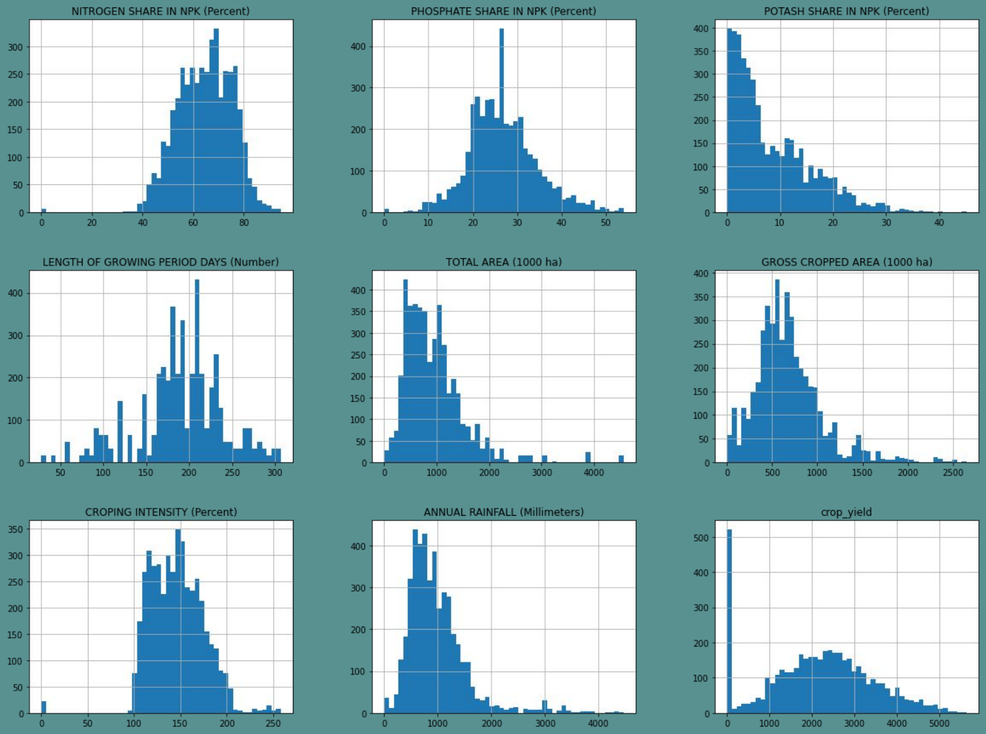
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Final Dataset

- Source
  - https://www.kaggle.com/datasets/anushkahedaoo/farming-factors
- Data Attributes

```
Data columns (total 10 columns):
 #   Column                                 Non-Null Count  Dtype  
---  ------                                 --------------  -----  
 0   NITROGEN SHARE IN NPK (Percent)        4480 non-null   float64
 1   PHOSPHATE SHARE IN NPK (Percent)       4480 non-null   float64
 2   POTASH SHARE IN NPK (Percent)          4480 non-null   float64
 3   LENGTH OF GROWING PERIOD DAYS (Number) 4480 non-null   float64
 4   TOTAL AREA (1000 ha)                   4480 non-null   float64
 5   GROSS CROPPED AREA (1000 ha)           4480 non-null   float64
 6   CROPING INTENSITY (Percent)            4480 non-null   float64
 7   ANNUAL RAINFALL (Millimeters)          4480 non-null   float64
 8   crop_name                              4480 non-null   object 
 9   crop_yield                             4480 non-null   float64
```

Final Dataset cont.

	NITROGEN SHARE IN NPK (Percent)	PHOSPHATE SHARE IN NPK (Percent)	POTASH SHARE IN NPK (Percent)	LENGTH OF GROWING PERIOD DAYS (Number)	TOTAL AREA (1000 ha)	GROSS CROPPED AREA (1000 ha)	CROPING INTENSITY (Percent)	ANNUAL RAINFALL (Millimeters)	crop_name	crop_yield
0	54.47	31.79	13.75	164.0	1988.17	1405.63	136.65	1149.7	RICE YIELD (Kg per ha)	1695.77
1	53.56	33.69	12.75	164.0	1988.17	1409.94	136.83	1282.3	RICE YIELD (Kg per ha)	1756.23
2	58.43	31.47	10.11	164.0	1988.17	1415.98	137.60	1092.1	RICE YIELD (Kg per ha)	1900.97

Methods

- We partitioned our data into train and test, with the train set accounting for 80 percent of the data
- Preprocessing data
  - Pipelines
  - One hot encoding
- Linear Regression Model
- Random Forest Regressor
- Decision Tree Regressor
- Support Vector Regression

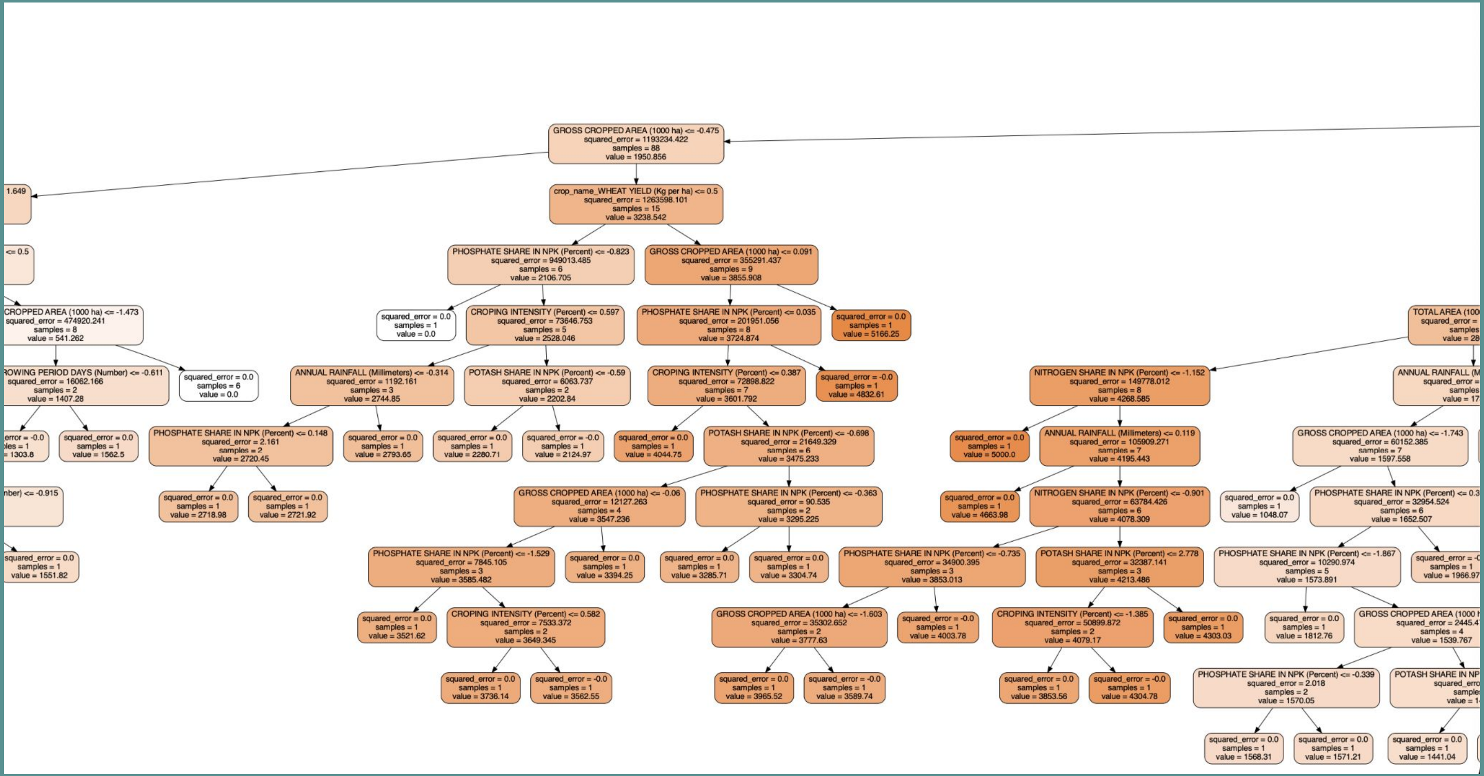
Initial Results

Models

Performance Metrics

	Linear Regression	Decision Tree	Random Forest	Support Vector Regression
RMSE	1173.88	850.207	612.02	814.228
MAE	933.955	562.774	443.38	612.915
R^2	0.144	0.551	0.767	0.588

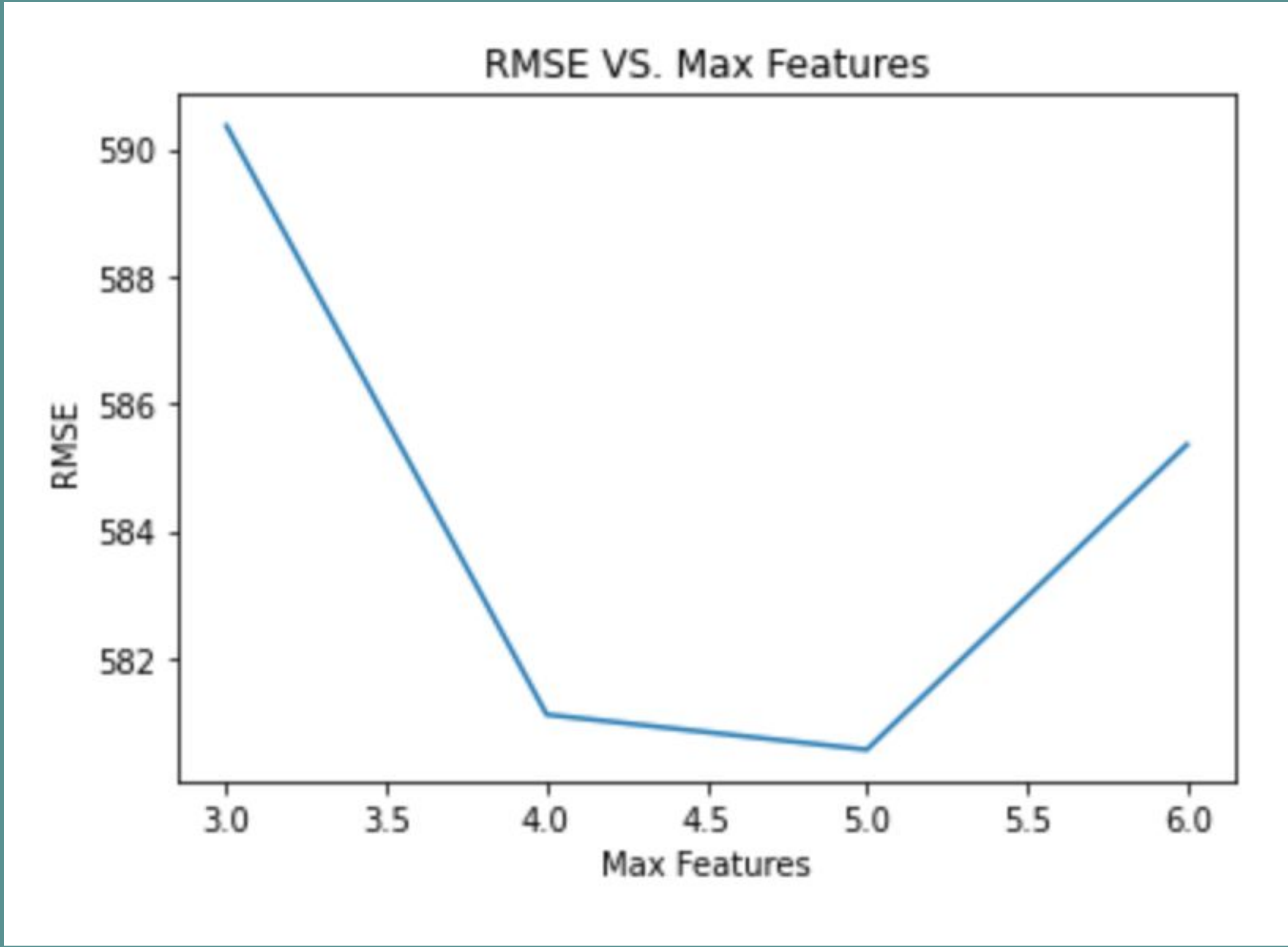
Initial Results



Hyperparameter Tuning (Random Forest)

Iteration No.

		1st (5-fold CV)	2nd (5-fold CV)	3rd (5-fold CV)	4th (5-fold CV)	5th (10-fold CV)
Hyperparameters	**bootstrap**	False	False	False	False	False
	max_depth	None	None	8	None	None
	max_features	3	3	5	6	5
	n_estimators	5000	1250	1000	1250	750
RMSE	**RMSE (train)**	604.725	605.019	712.257	596.738	580.562
	RMSE (test)	592.591	593.058	709.209	582.429	581.173


RMSE VS. Max Features
590
588
586
584
582
RMSE
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Max Features

Variable Importance (Random Forest)

Random Forest Regressor
(Initial)

0.1915	Cropping Intensity (%)
0.1606	Potassium Share in NPK (%)
0.1223	Total Area (1000 ha)
0.1172	Gross Cropped Area (1000 ha)
0.1117	Length of Growing Period (Days)
0.0736	Annual Rainfall (mm)
0.0726	Phosphate Share in NPK (%)
0.0642	Nitrogen Share in NPK (%)

Random Forest Regressor
(Final)

0.1785	Cropping Intensity (%)
0.1495	Potassium Share in NPK (%)
0.1302	Total Area (1000 ha)
0.1231	Length of Growing Period (Days)
0.1203	Gross Cropped Area (1000 ha)
0.0812	Annual Rainfall (mm)
0.0748	Phosphate Share in NPK (%)
0.0653	Nitrogen Share in NPK (%)

Conclusion

- Standardized RMSE: Test = 0.1047, Train = 0.1028
  - Predicting crop yield with minimal error (~10%)
- India is the second largest producer of rice and wheat
  - Critical staple crops to world food supply
- Ensure a maintained supply even as the environment changes
- Feature importance:
  - Future work with potassium

Future Work

- **Limitations to Address**
  - Little information about how the data was collected, possibility of measurement errors
  - Lack of description for some column names in the dataset
  - Relatively small dataset
- **Further Analysis**
  - Testing additional crop types beyond rice and wheat
  - Including additional environmental conditions, such as temperature, soil properties, and climate
  - Investigating whether the findings can be applied to other parts of the world
  - Exploring trends over time
  - Understanding the effect of droughts or diseases that may have inadvertently affected the crop yield