COS60008 Introduction to Data Science

Analysing Agricultural Yield through Data Analytics

Kan Asvasena

Submission Date: 23 June 2023

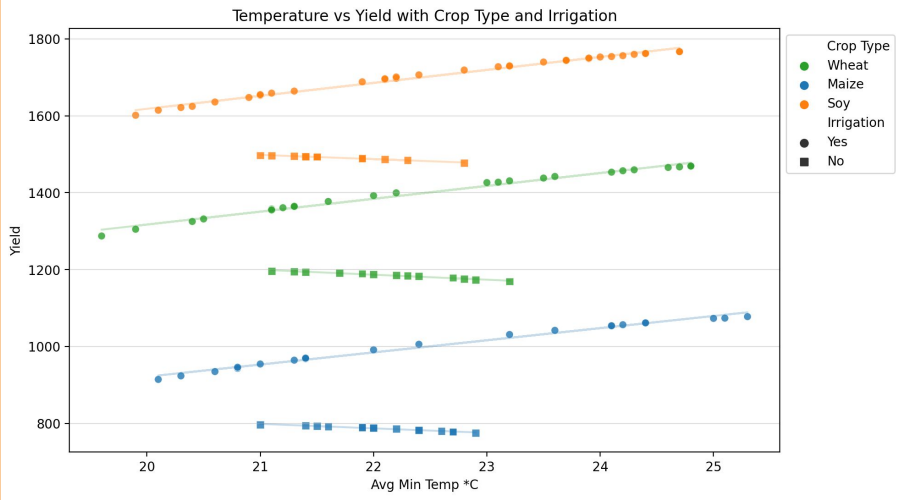
1. Data Preparation

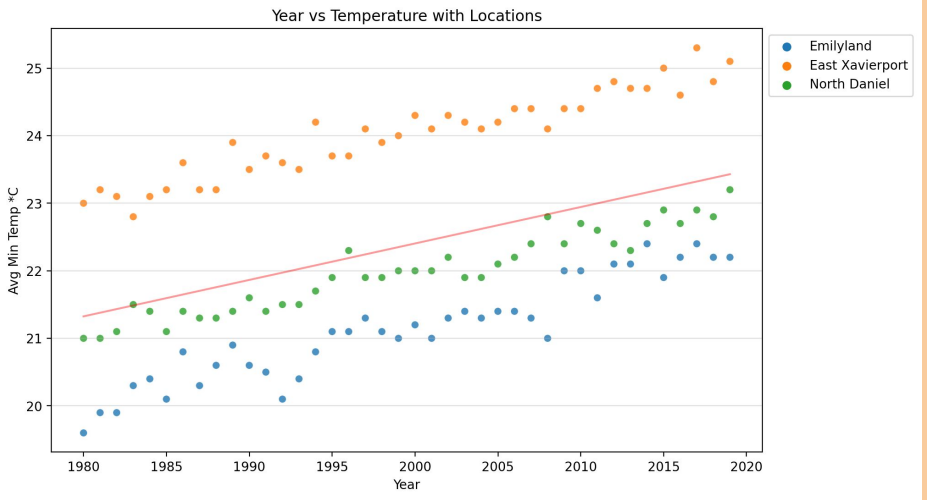
- Data Checking
- Data Cleaning
- Data Merging

	column	dtype	range/unique
0	index	int64	0 - 119
1	Year	int64	1980 - 2019
2	Location	object	['Emilyland' 'East Xavierport' 'North Daniel']
3	Avg Min Temp *C	float64	19.6 - 25.3
4	Yr Rain mm	float64	591.4415399999999 - 2097.95
5	Heat Wave	object	['N' '0' 'Y' 'no' '1']
6	Dry Spell	object	['N' 'n' 'No' 'Y' '0' 'no']
7	Cold Wave	object	['Y' 'N' 'yes' 'No' '1' '0']
8	Wet Spell	object	['N' 'no' 'Y' 'Yes']
9	Irrigation	object	['Y' 'N' 'y' '1' 'no']
10	Crop Type	object	['Wheat' 'Maize' 'Soy']
11	Crop Damage	float64	0.0 - 1.0
12	Yield	float64	775.62 - 1766.74
13	Observer	object	['Ryan Stewart' 'Tracie Bell' 'Thomas Miller' 'Sara Nguyen' 'Tammy Calderon' 'Amanda Reed' 'Kevin Henry' 'Emily Bernard' 'Noah Maldonado' 'James Todd' 'Diana Jones' 'Patricia Lopez' 'Jennifer Patel' 'Karina Newman' 'William Wright' 'John Martinez' 'Jessica Hodges' 'Jonathan Gregory' 'Martha Johnson' 'Sandra Thomas' 'Robert George' 'Alexander Johnson' 'Kelly George' 'Nicole Wall' 'Natalie Branch' 'Martin Rivera' 'Mary Torres' 'Tiffany Decker' 'Stephen Nelson' 'Joe James' 'Richard Gardner' 'Christina Nichols' 'Sylvia Hanson' 'Mary Day' 'Samantha Murphy' 'Michael Norris' 'Stephen Tran DDS' 'James Schneider' 'Tammy Carlson' 'William Keller' 'Diana Elliott' 'Renee Moore' 'Bonnie Sanders' 'Elizabeth Castro' 'Misty

2. Key Findings from Data Exploration (1)

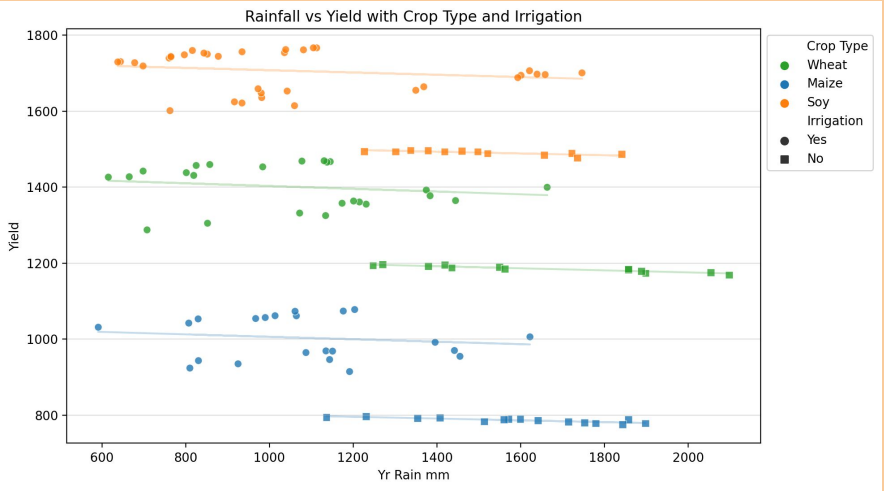
- Temperature vs Yield
  - Rising temperatures positively affect yield in irrigated areas but negatively in non-irrigated one.
  - Temperature has been increasing every year, so it positively impact on increasing yields in irrigated locations.

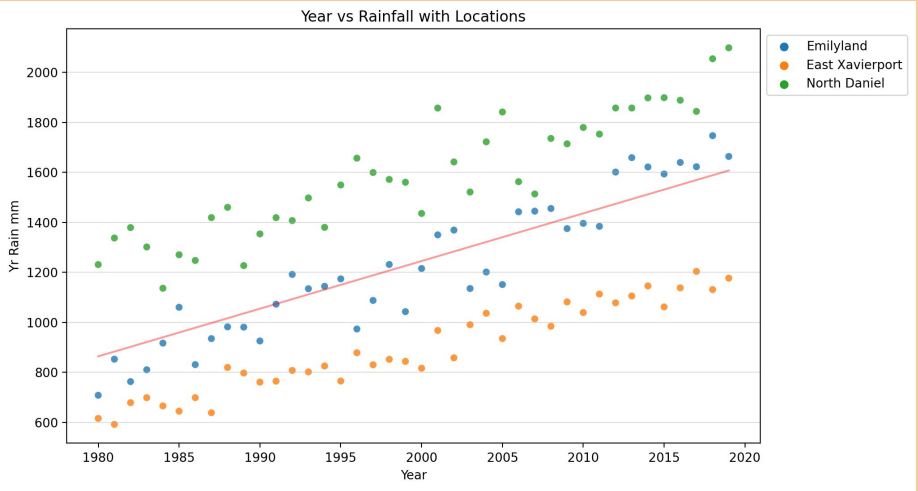




2. Key Findings from Data Exploration (2)

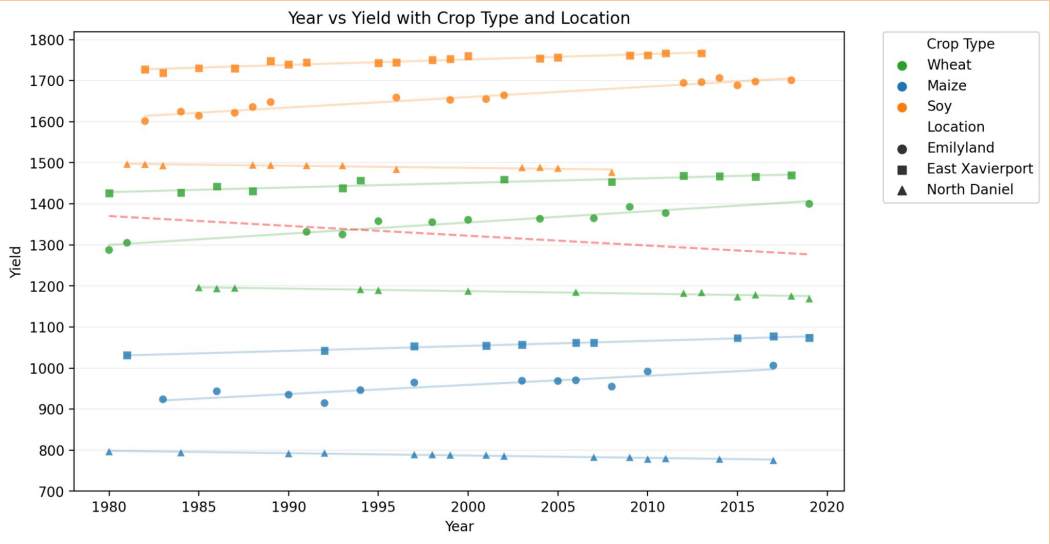
- Rainfall vs Yield
  - As rainfall increased, yield slightly decreased.
  - Rainfall has been increasing every year, so it has a minor negative impact on the overall yield.





2. Key Findings from Data Exploration (3)

- Overall, yield in irrigated locations, Emilyland and East Xavierport, has increased each year.
- Whereas, non-irrigated location, North Daniel, slightly decreased.
- The average trend line shows a decline due to reduced production of the high-yield crop, soy.



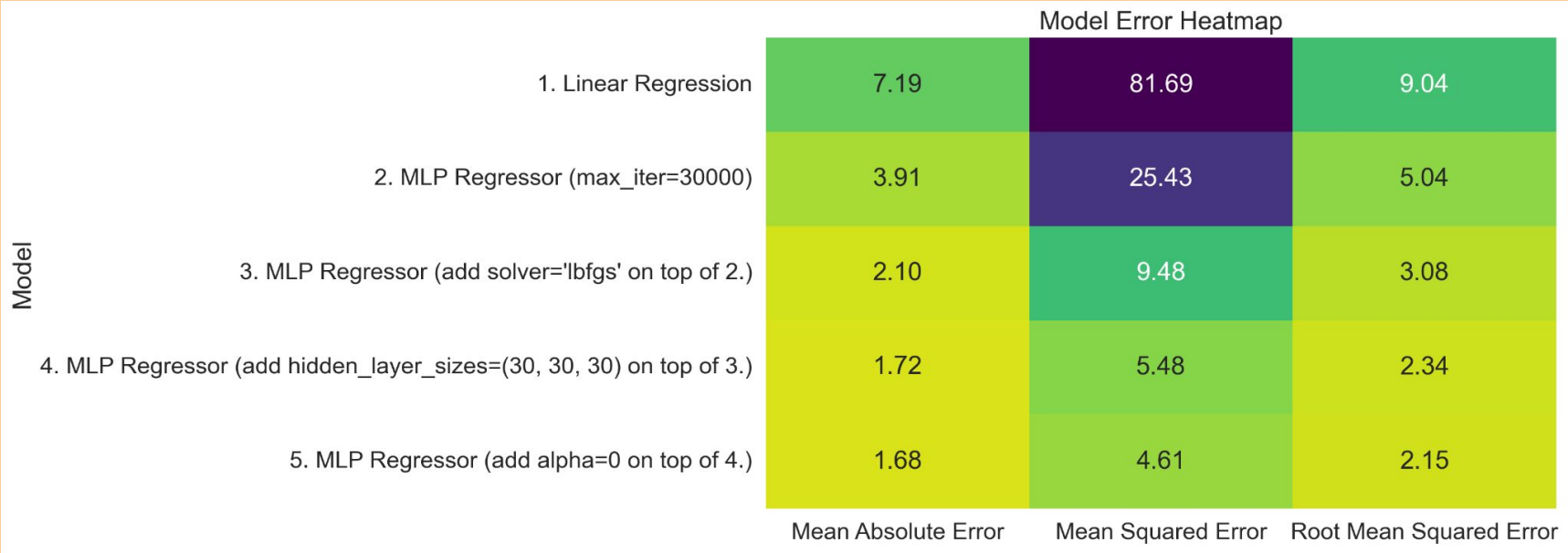
3. Data Modeling

- To predict the yield
- 5 Models
  - 1 Linear Regression
  - 4 Multi-layer Perceptron (MLP) Regression
    - added hyperparametres incrementally on top of the previous one

Model	Hyperparametre
1. Linear Regression	default
2. MLP Regressor	max_iter=30000
3. MLP Regressor	max_iter=30000, solver='lbfgs'
4. MLP Regressor	max_iter=30000, solver='lbfgs', hidden_layer_sizes=(30, 30, 30)
5. MLP Regressor	max_iter=30000, solver='lbfgs', hidden_layer_sizes=(30, 30, 30), alpha=0

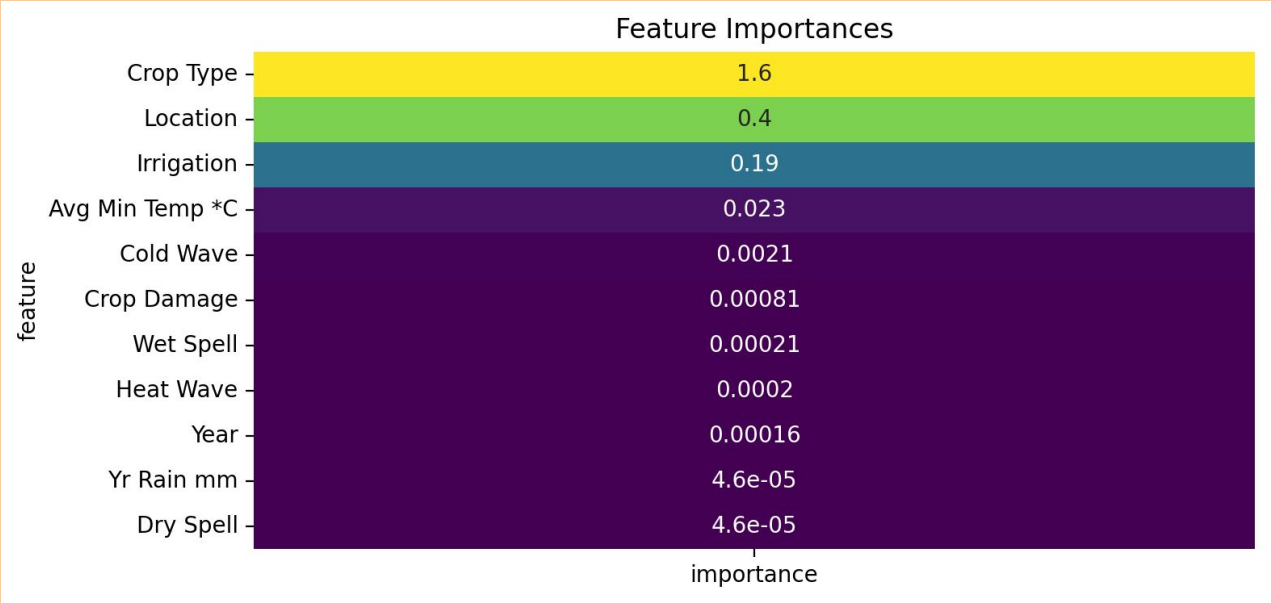
4. Model Evaluation

- **The best model** is the final MLP Regressor with max_iter=30000, solver='lbfgs', hidden_layer_sizes=(30, 30, 30), and alpha=0



Featured Importances

- The Permutation Importance method is used to determine the importance of each feature on the best performing model.



Thank you