ECE 592 – Topics in Data Science HW06

Ashwini Muralidharan (200483207)
Purushothaman Saravanan (200483316)
Shahil Manoj Dhotre (200475496)

Introduction :

Crop yield prediction remains one of the most challenging tasks in agriculture as it affects decision making at global, regional, and field levels. The prediction of crop yield is based on several factors such as soil, meteorological, environmental, and crop parameters. Due to the colossal effect agriculture has on the economy of the country, data-driven prediction models need to be developed. Against this background, machine learning (ML) models were engineered after processing agricultural data and Regression models were deployed to predict the yield. The developed models were then compared with each other to obtain the best performing model.

Dataset :

The first input dataset was the yield data of crops from all the states in the United States of America, obtained from the FAOSTAT database. This project focuses on 4 major crops of this region, namely - wheat, rice, bananas and avocados for the years 1961 to 2020. Relevant information of fertilizer usage and export quantity of the 4 crops were also extracted from the same source and combined with each crop. This information was matched with the second dataset pertaining to the climatic conditions of each year, obtained from Climate Change Knowledge portal. The important climatic parameters were chosen to be Average Temperature and Precipitation. The final dataset contained the features listed in Table 1.

Feature	Unit
Precipitation	mm
Average Temperature	°C
Export Quantity	tonnes
Fertilizer used	tonnes

Table 1: Features considered

The independent variables i.e., features are listed in Table 1, and the dependent variable to be predicted is the crop yield. Figure 1 shows the first few rows of the merged dataset.

	Crop	Year	Yield	Temperature (Avg)	Precipitation	Export Quantity	Fertilizer Usage
235	Wheat	2016	35408	10.62	746.82	24041586.0	20091404.3
236	Wheat	2017	31175	10.21	777.23	27299214.0	20145604.7
237	Wheat	2018	32005	9.87	822.17	22499006.0	20018597.5
238	Wheat	2019	34746	9.72	834.45	27068607.0	19950557.9
239	Wheat	2020	33415	10.00	737.57	26131626.0	19899437.6

Figure 1: Dataset sample

Data Pre-processing :

The data was pre-processed to prepare it for the regression analysis. The empty/missing values were dealt with using linear interpolation method. The data was further rescaled using normalization techniques. The feature column of ‘Crop’ was One Hot Encoded to deal with categorical data. The final processed dataset was split into training and testing set in the ratio of 80:20 to deploy the regression models.

	Temperature (Avg)	Precipitation	Export Quantity	Fertilizer Usage	Crop_Avocados	Crop_Bananas	Crop_Rice	Crop_Wheat
0	-1.048378	0.236993	-0.578619	-3.149149	1	0	0	0
1	-0.547368	-1.099606	-0.578619	-2.905647	1	0	0	0
2	-0.270949	-2.127306	-0.578619	-2.636126	1	0	0	0
3	-1.307521	-0.496205	-0.578619	-2.448389	1	0	0	0
4	-1.031102	-0.273693	-0.578619	-2.144140	1	0	0	0

Figure 2: Data after one hot encoding

Data Visualization :

The distribution of the 4 crops across the USA from the year 1960 to 2020 is illustrated in Figure 3.

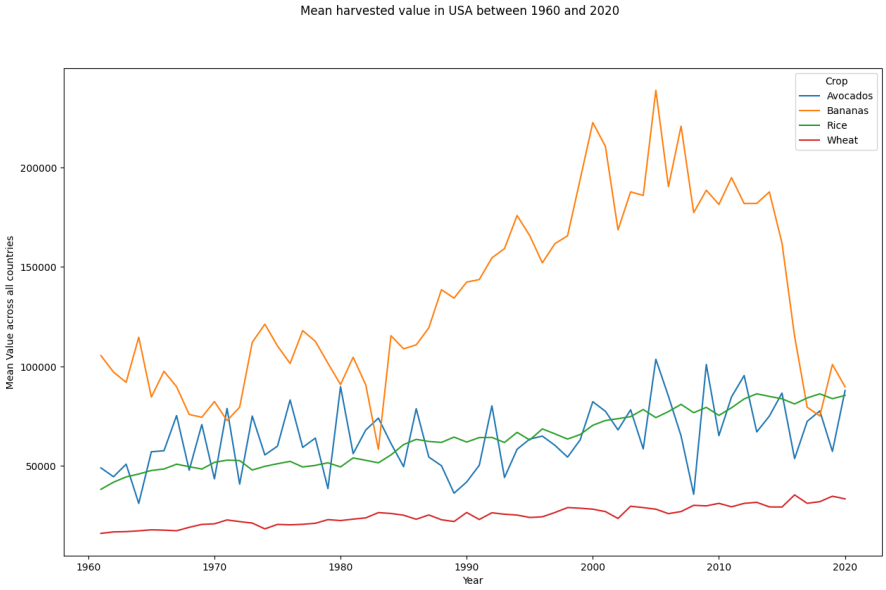

Figure 3: Mean harvested value vs Years

The correlation between the features is calculated and tabulated as shown below in Figure 4. It is also depicted in the form of a heatmap for clearer visualization in Figure 5.

	Yield	Temperature (Avg)	Precipitation	Export Quantity	Fertilizer Usage
Yield	1.000000	0.206763	0.019262	-0.569596	0.209991
Temperature (Avg)	0.206763	1.000000	0.223916	-0.015737	0.509243
Precipitation	0.019262	0.223916	1.000000	0.013164	0.324463
Export Quantity	-0.569596	-0.015737	0.013164	1.000000	0.047415
Fertilizer Usage	0.209991	0.509243	0.324463	0.047415	1.000000

Figure 4: Correlation coefficients for the feature variables

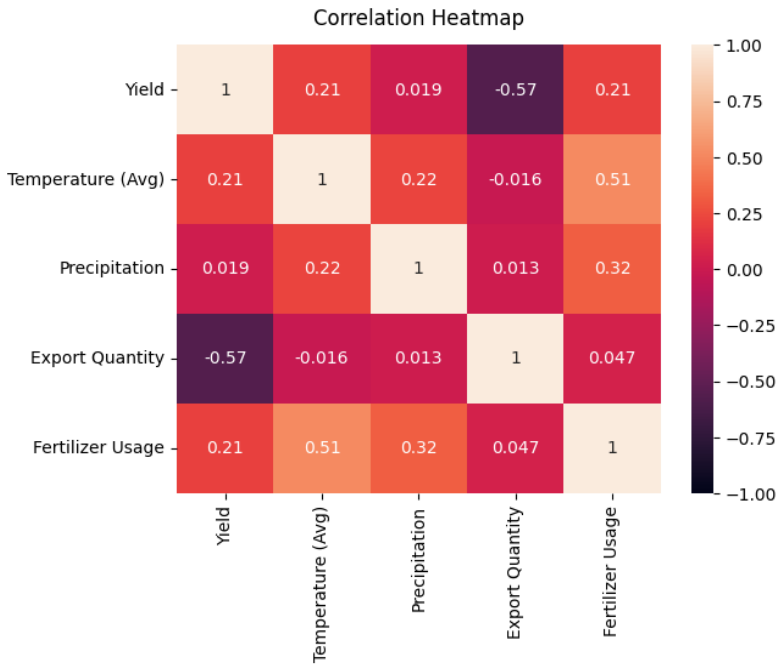

Figure 5: Heatmap of the correlation coefficients for the feature variables

According to the Heatmap in Figure 4:
Temperature (Avg) is moderately positive correlated with yield
Fertilizer Usage is moderately positive correlated with yield
Precipitation is low positively correlated with yield
Export Quantity is highly negatively correlated with yield.

ML models :

In order to predict the crop yield for each crop, regression models were deployed, whose task is the prediction of the dependent variable with the help of other independent variables. Keeping 'Crop' as the ground truth value and the feature columns referenced in Table 1, five regression models, namely - Multivariate Linear Regression, Shrinkage methods (Ridge and Lasso regression), k-nearest neighbors, Decision Tree and Artificial Neural Networks were trained.

- **Metrics** :
  The performance of regression models were assessed using four important metrics, namely - Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R2. The results of the ML models using the above metrics are tabulated below.

Regressor	MAE	MSE	RMSE	R2
Multivariate Linear regression	14735.19	413785184.80	20341.71	0.83
Ridge regression	14790.58	420434163.38	20504.49	0.82
Lasso regression	14735.21	413786300.68	20341.73	0.83
Decision Tree regression	10626.41	248188971.79	15754.01	0.89
K-nearest Neighbors	12449.21	429741854.49	20730.21	0.82
Artificial Neural Network	11406.93	365213980.0	19110.57	0.85

Table 2: Regression Results

Against the optimum performance standards for this task, after observation and evaluation of the metrics of the regressors, Decision Tree regressor exhibited best performance.

The analysis thus retrieved can be used as a basic prototype by the agricultural industry for further research by using better data, advanced pre-processing techniques and machine learning models.