YIELD ESTIMATION

- Problem Statement :

Estimating Yield Of Particular Crop.

- Objective :

The objective is to enhance accuracy by leveraging various features through the implementation of machine learning algorithms and employing feature selection techniques.

❑Data Collection:

- The data exists in two datasets(GP based and Location based).
- Data points were replaced in the column based on Gram Panchayat and Tehsil Block.
- The columns that are similar in both GP and Location based:
  - NDVI (Normalized Difference Vegetation Index)
  - LAI (Leaf Area Index)
  - LAI (Leaf Area Index)
  - FAPAR (Fraction of Absorbed Photosynthetically Active Radiation)
  - FCover (Fractional Crop Cover)

❑ Data Exploration:

- Measure Of Central Tendency

mean = 33.5

median=20.9

If , mean > median i,e.. Outliers On Higher Side

- <u>Measure Of Dispersion:</u>

Variance = 1536.0992

Standard Deviation = 39.1931

High Standard Deviation indicates that the values are spread over a wider range.

Here , standard deviation > mean then predictability chances are high.

It falls under positive skewness . X directly proportional to y.

❑DATA CLEANING:

- Replaced “Nodata” points with NaN using Numpy.
- For continuous data:Imputed NaN values using corresponding summary statistics such as mean, median, and mode.
- Converted the columns to their appropriate data types.
- For Categorical data:"Performed one-hot encoding for the features 'Any Damage,' 'Weeds,' and 'Crop Condition'.
- Identified outliers using Isolation Forest and subsequently removed them.
- Final dataset contains 120 samples and 90 features.

['Experimental weight', 'Sowing_Area', 'SWC_Latitu', 'SWC_Longit', 'rf2fnjul22', 'rf1fnaug22', 'rf2fnaug22', 'rf1fnsep22', 'rf2fnsep22', 'rf1fnoct22', 'rf2fnoct22', 'Tmax_2fnjul22', 'Tmax_1fnaug22', 'Tmax_2fnaug22', 'Tmax_1fnsep22', 'Tmax_2fnsep22', 'Tmax_1fnoct22', 'Tmax_2fnoct22', 'Tmin_2fnjul22', 'Tmin_1fnaug22', 'Tmin_2fnaug22', 'Tmin_1fnsep22', 'Tmin_2fnsep22', 'Tmin_1fnoct22', 'Tmin_2fnoct22', 'rh_max_2fnjul22', 'rh_max_1fnaug22', 'rh_max_2fnaug22', 'rh_max_1fnsep22', 'rh_max_2fnsep22', 'rh_max_1fnoct22', 'rh_max_2fnoct22', 'rh_min_2fnjul22', 'rh_min_1fnaug22', 'rh_min_2fnaug22', 'rh_min_1fnsep22', 'rh_min_2fnsep22', 'rh_min_1fnoct22', 'rh_min_2fnoct22', 'RVIJul2fn', 'RVIAug1fn', 'RVIAug2fn', 'RVISep1fn', 'RVISep2fn', 'RVIOct1fn', 'RVIOct2fn', 'NDVI_2fnjul22', 'NDVI_1fnaug22', 'NDVI_2fnaug22', 'NDVI_1fnsep22', 'NDVI_2fnsep22', 'NDVI_1fnoct22', 'NDVI_2fnoct22', 'LAI_2fnjul22', 'LAI_1fnaug22', 'LAI_2fnaug22', 'LAI_1fnsep22', 'LAI_2fnsep22', 'LAI_1fnoct22', 'LAI_2fnoct22', '2fnJul22_FCover', '1fnAug22_FCover', '2fnAug22_FCover', '1fnSep22_FCover', '2fnSep22_FCover', '1fnOct22_FCover', '2fnOct22_FCover', 'FAPAR_2fnJuly', 'FAPAR_1fnAug', 'FAPAR_2fnAug', 'FAPAR_1fnSept', 'FAPAR_2fnSept', 'FAPAR_1fnOct', 'FAPAR_2fnOct', '2fnJul22_DryMatter(Biomass)', '1fnAug22_DryMatter(Biomass)', '2fnAug22_DryMatter(Biomass)', '1fnSep22_DryMatter(Biomass)', '2fnSep22_DryMatter(Biomass)', '1fnOct22_DryMatter(Biomass)', '2fnOct22_DryMatter(Biomass)', 'Any_Damage_0', 'Any_Damage_1', 'Any_Damage_2', 'Weeds_0', 'Weeds_1', 'Weeds_2', 'Crop condition_0', 'Crop condition_1', 'Crop condition_2'],

❑ Here, Is the correlation matrix between independent variables (features) and the dependent variable(Target).

0 - 0.19 : very weak

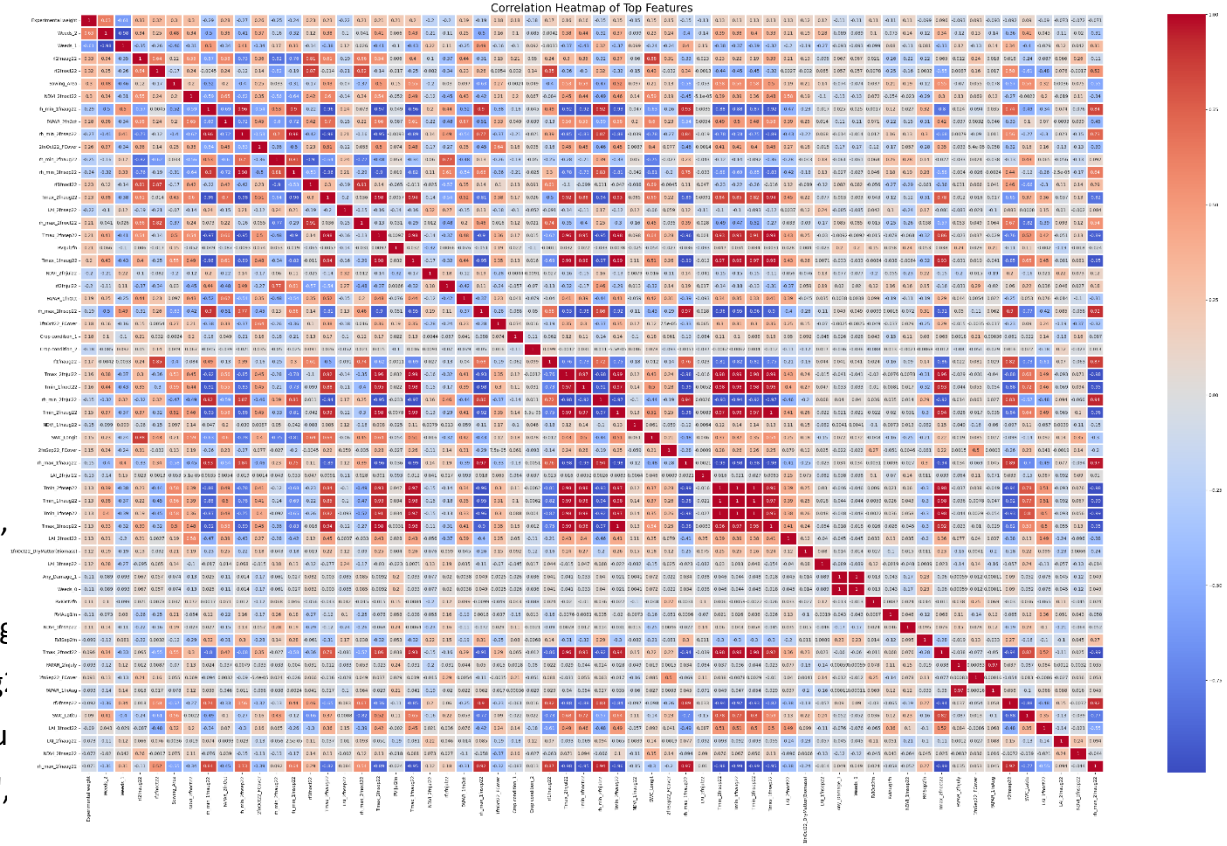
0.2 - 0.39 : weak

0.4 - 0.59 : moderate

0.6 – 0.79 : strong

0.81 - 1 : perfect

Vice versa for negative



Selected Features :

['Experimental weight', 'Weeds_2', 'Weeds_1', 'rf2fnaug22', 'rf2fnoct22',
'Sowing_Area', 'NDVI_2fnoct22', 'rh_min_1fnaug22', 'FAPAR_2fnOct',

- 'rh_min_2fnsep22', '2fnOct22_FCover', 'rh_min_2fnaug22',
- 'rh_min_1fnsep22', 'rf1fnoct22', 'Tmax_2fnaug22', 'LAI_2fnsep22',
- 'rh_max_2fnoct22', 'Tmax_2fnsep22', 'RVIJul2fn', 'Tmax_1fnaug22',
- 'NDVI_2fnjul22', 'rf2fnjul22', 'FAPAR_1fnOct', 'rh_max_1fnsep22',
- '1fnOct22_FCover', 'Crop condition_1', 'Crop condition_2', 'rf1fnaug
- 'Tmax_2fnjul22', 'Tmin_1fnoct22', 'rh_min_2fnjul22', 'Tmin_2fnaug
- 'NDVI_1fnaug22', 'SWC_Longit', '2fnSep22_FCover', 'rh_max_1fnau
- 'LAI_2fnjul22', 'Tmin_2fnsep22', 'Tmin_1fnaug22', 'Tmin_1fnsep22',
- 'Tmax_1fnsep22', 'LAI_2fnoct22', '1fnOct22_DryMatter(Biomass)',
- 'LAI_1fnsep22', 'Any_Damage_1', 'Weeds_0', 'RVIOct2fn', 'RVIAug1fn',
- 'NDVI_1fnsep22', 'RVISep2fn', 'Tmax_2fnoct22', 'FAPAR_2fnJuly',
- '1fnSep22_FCover', 'FAPAR_1fnAug', 'rf2fnsep22', 'SWC_Latitu', 'LAI_1fnoct22', 'LAI_2fnaug22', 'NDVI_2fnsep22', 'rh_max_2fnaug22']

❑ Data Preprocessing :

OLS Selected Features Based on significance value(0.05) :

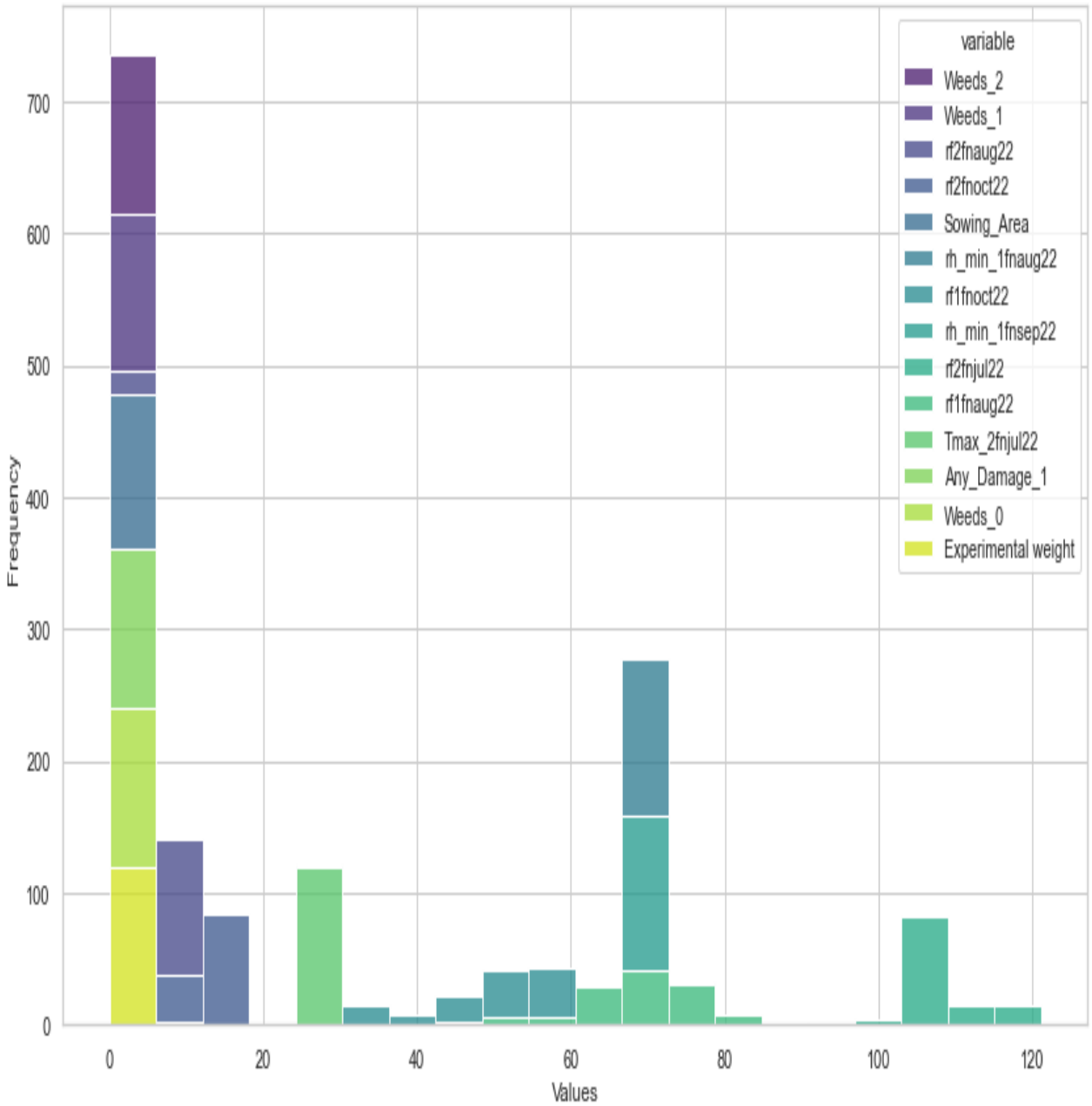
['Weeds_2', 'Weeds_1', 'rf2fnoct22', 'rf2fnaug22', 'Sowing_Area','rf1fnoct22', 'rf1fnaug22', 'RVIJul2fn', 'rh_min_1fnaug22','rh_min_1fnsep22', 'rf2fnjul22', 'Any_Damage_1', 'Any_Damage_2','Weeds_0', 'Tmax_2fnjul22', 'NDVI_2fnsep22', 'Experimental weight']

Feature Scaling:

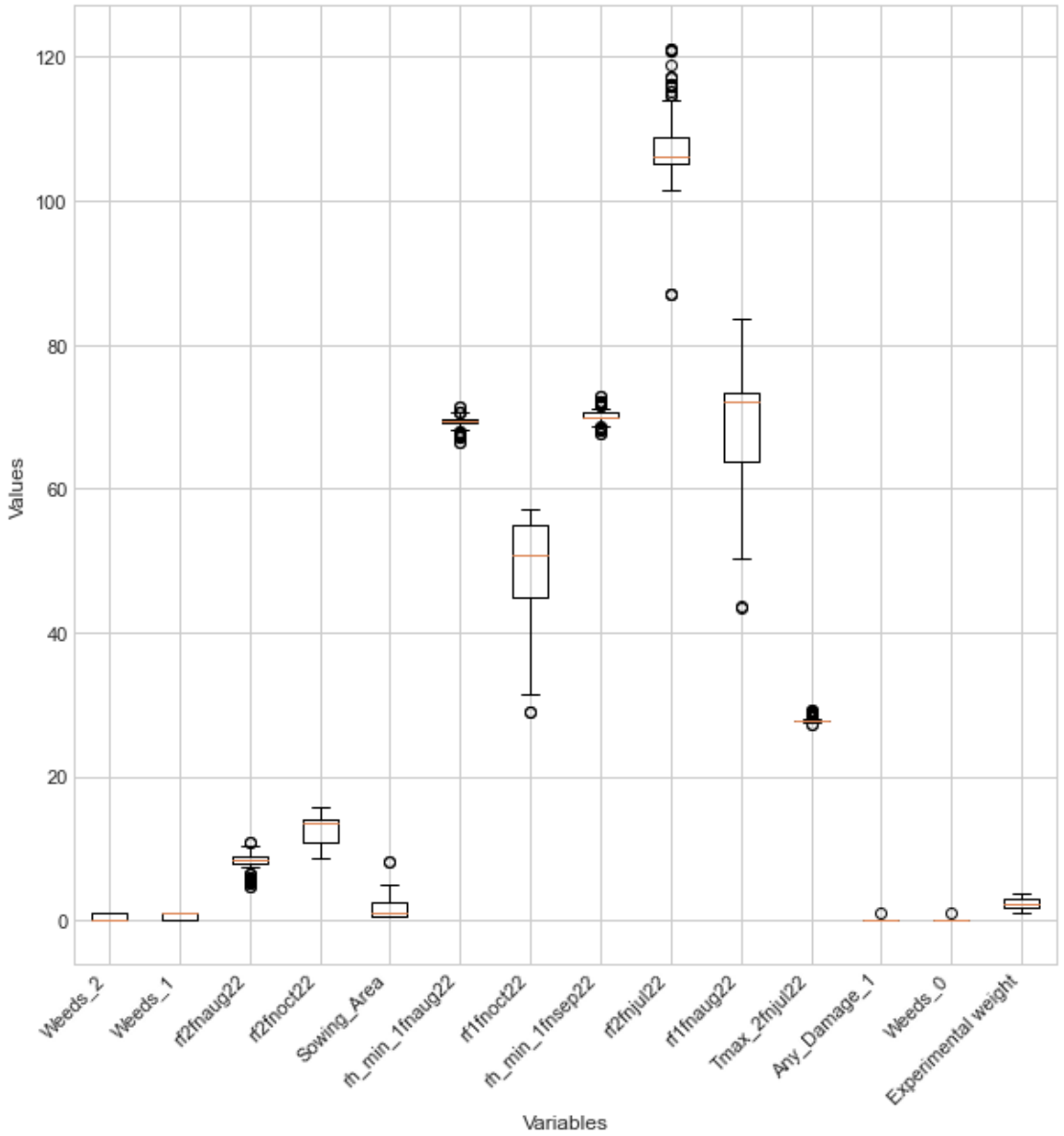
1.Standardization(Z-score normalization):

Standardized the features selected through Recursive Feature Elimination (RFE) to have a mean of 0 and a standard deviation of 1.

Front View Of Data



Back View Of Data



❑ Model Selection and Model Training:

- Train-Test-Split: Dataset that contains both input features (X) and corresponding target values or labels (y). Splitted into two disjoint sets(Training set and Testing set). Test-size 0.02(80-20)

XTRAIN (96, 16)

XTEST (24, 16)

YTRAIN (96,1)

YTEST (24,1)

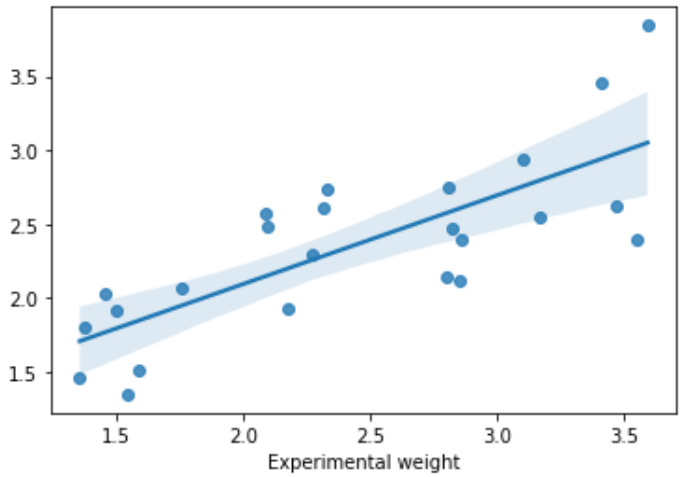
Linear regressor :

Test Accuracy : 0.5780

Train Accuracy : 0.6015

MSE(Mean Squared Error) : 0.2241

MAE(Mean Absolute Error) : 0.3871



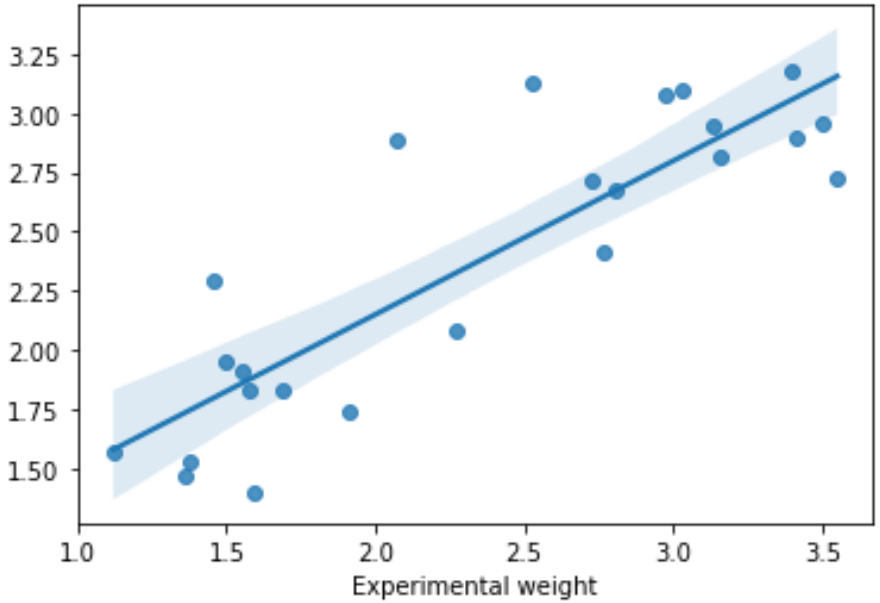
Random Forest Regressor :

Test Accuracy : 0.7319

Train Accuracy : 0.8842

MSE(Mean Squared Error) : 0.1690

MAE(Mean Absolute Error) : 0.3332



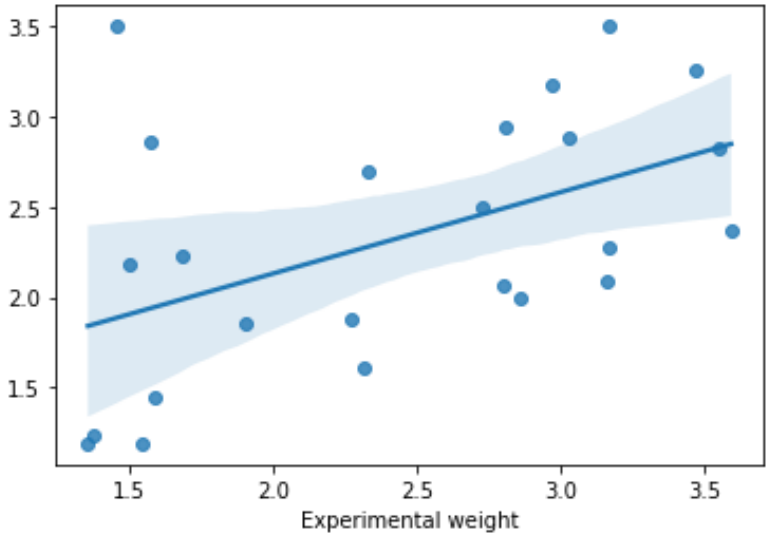
Decision Tree Regressor :

Test Accuracy : 0.042

Train Accuracy : 0.9698

MSE(Mean Squared Error) : 0.5441

MAE(Mean Absolute Error) : 0.5685



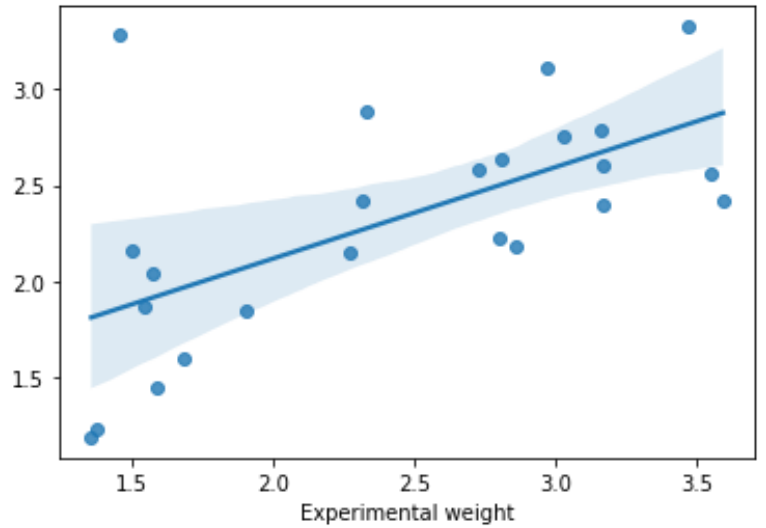
XG Boost Regressor :

Test Accuracy : 0.3506

Train Accuracy : 0.9698

MSE(Mean Squared Error) : 0.3690

MAE(Mean Absolute Error) : 0.4447



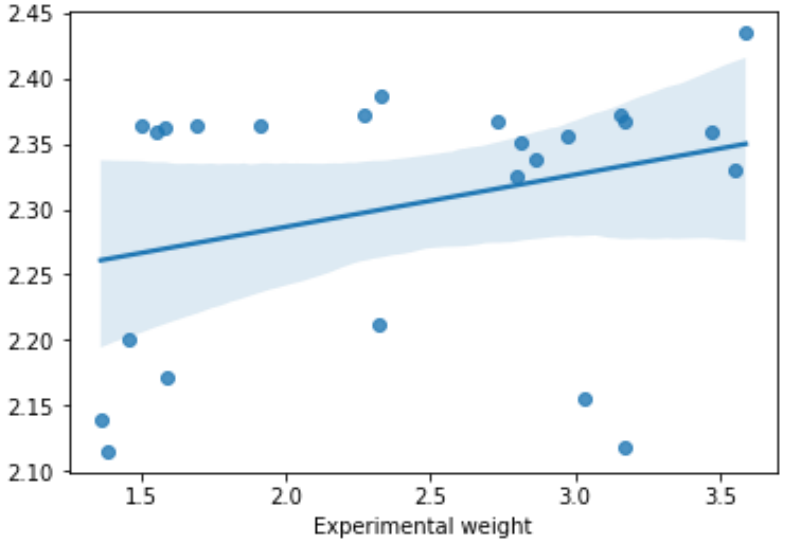
SVM Regressor :

Test Accuracy : 0.0363

Train Accuracy : 0.0622

MSE(Mean Squared Error) : 0.5476

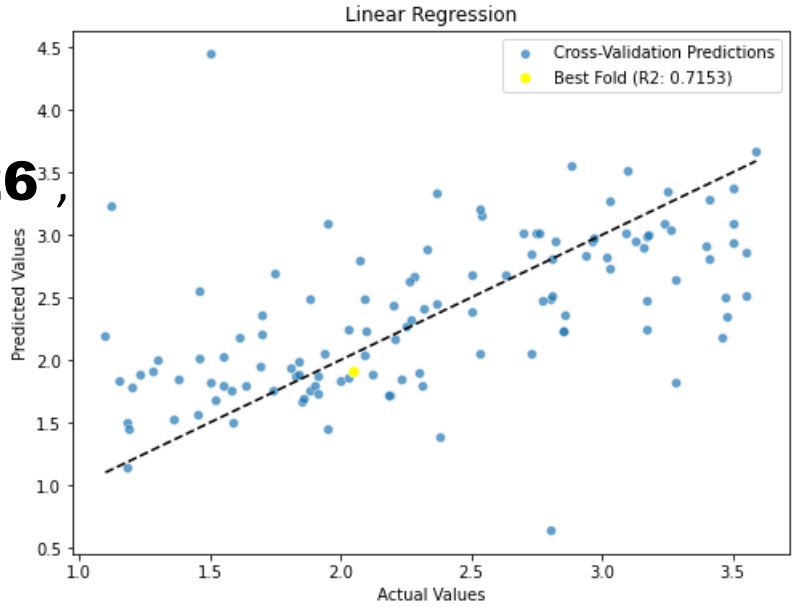
MAE(Mean Absolute Error) : 0.6718



Conducting algorithms with cross-validation :

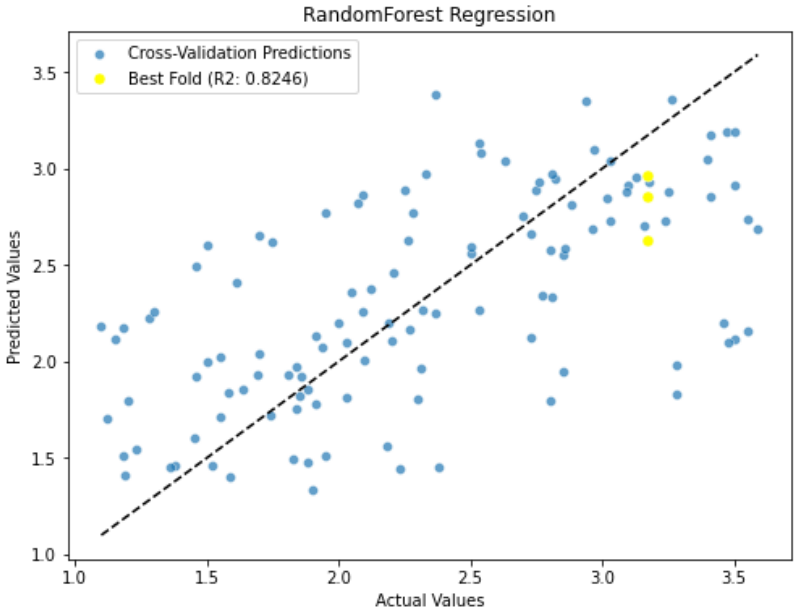
Linear Regression :

Cross-Validation Scores : [0.587754 , **-0.75719526** ,
0.65047816 ,0.2306657 ,-0.23337346 -1.9708962,
0.31016553 -0.12654222 ,**0.71533678**]



Random Forest Regressor :

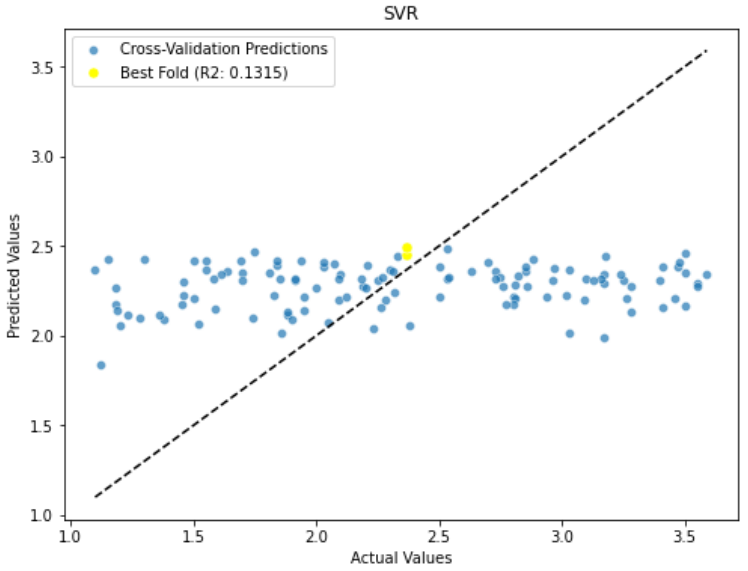
Cross-Validation Scores : [0.39061098 , 0.35629638 ,
0.05340942 0.09478889 , -0.51794765 ,
-0.07119532 , **0.82458419**]

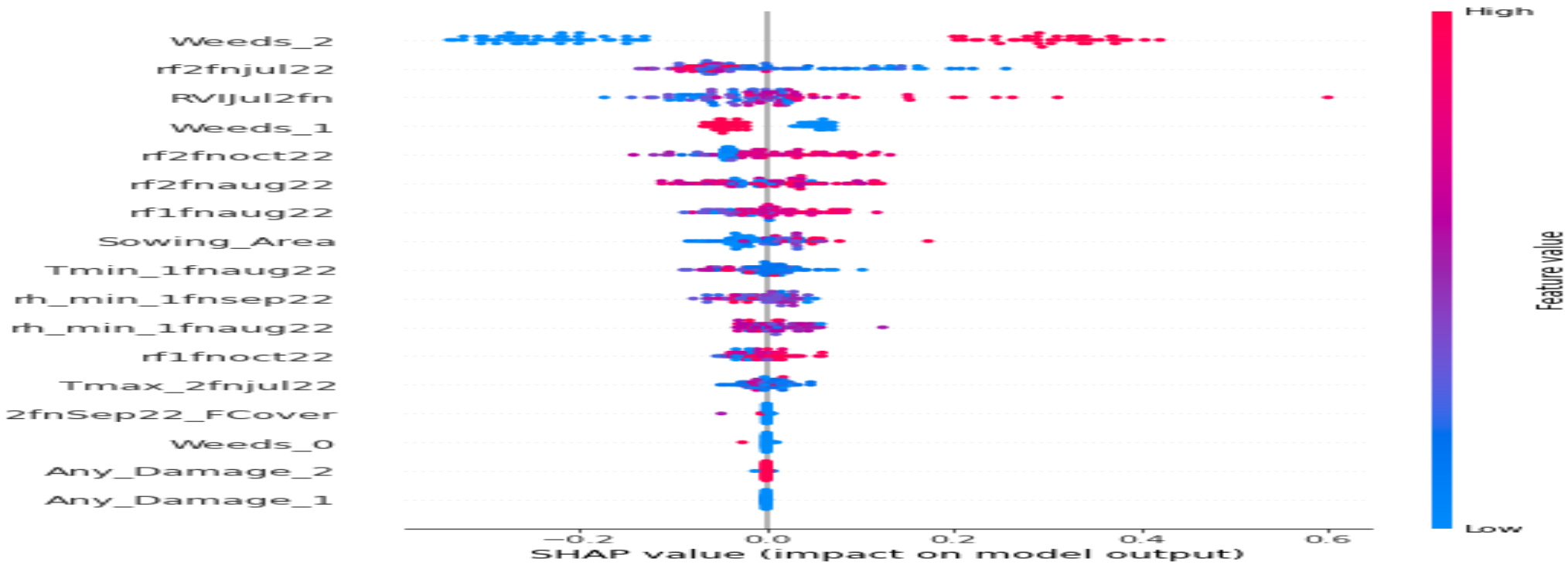


Decision Tree Regressor : [-0.4294819 , -0.39210673 , -0.5090972 , **-0.83848352** , -1.45319263 , -0.30321498 , **0.65411139**]

SVM Regressor :

[**0.13147365** , -0.00984736 , **-0.23161514** , -0.03874791 ,-1.20441983 , -0.22837572 , 0.07955813]





Most impactful features: ['Weeds_2', 'rf2fnjul22', 'RVIJul2fn', 'Weeds_1', 'rf2fnoct22','rf2fnaug22', 'rf1fnaug22', 'Sowing_Area', 'Tmin_1fnaug22', 'rh_min_1fnsep22', 'rh_min_1fnaug22', 'rf1fnoct22', 'Tmax_2fnjul22','2fnSep22_FCover', 'Weeds_0', 'Any_Damage_2', 'Any_Damage_1']

Conclusion : I achieved 82% accuracy using Random Forest, employing various feature selection techniques along with cross-validation.