CROP YIELD PREDICTION IN INDIA

Predicting yield helps the state to get an estimate of the crop in a certain year to control the price rates.This model focuses on predicting the crop yield in advance by analyzing factors like location, season, and crop type through machine learning techniques on previously collected datasets.

Dataset Overview

The dataset contains the following columns:

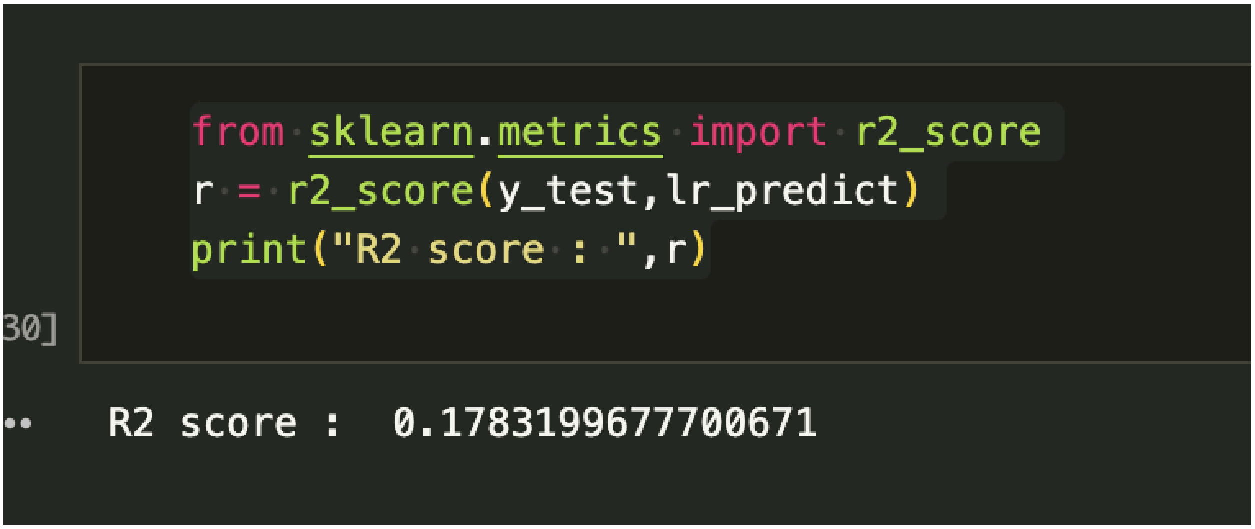
1. Crop_Year: The year of crop production.
2. State_Name: The state where the crop was produced.
3. District_Name: The district within the state.
4. Season: The season during which the crop was grown.
5. Crop: The type of crop.
6. Area: The area under cultivation (in hectares).
7. Production: The total production of the crop (in tons).
8. The target variable is Yield, which is calculated as Production / Area.

Preprocessing Steps and Observations

1. Handling Missing Values
    a. Observation: The dataset had missing values in the Area and Production columns.
2. Dropping Rows with Missing Critical Data
    a. Observation: Some rows had missing values in the Crop and Season columns.
3. Data Type Conversion
    a. Observation: The data types of some columns were not optimal for analysis.
4. Removing Duplicates
    a. Observation: The dataset contained duplicate rows.
5. Standardizing Column Names
    a. Observation: Column names had inconsistent casing.
6. Cleaning Categorical Data
    a. Observation: The Season column had leading/trailing spaces.
7. Calculating Yield
    a. Observation: The target variable Yield was not directly available in the dataset.
8. Normalizing Numerical Features
    a. Observation: Numerical features (Crop_Year, Area, Production) had different scales.
9. Filtering Invalid Data
    a. Observation: Some rows had zero or negative values for Production.
10. Saving Preprocessed Data
    a. Action: The preprocessed dataset was saved to a new CSV file (preprocessed_crop_production.csv).

Observations on Linear Regression

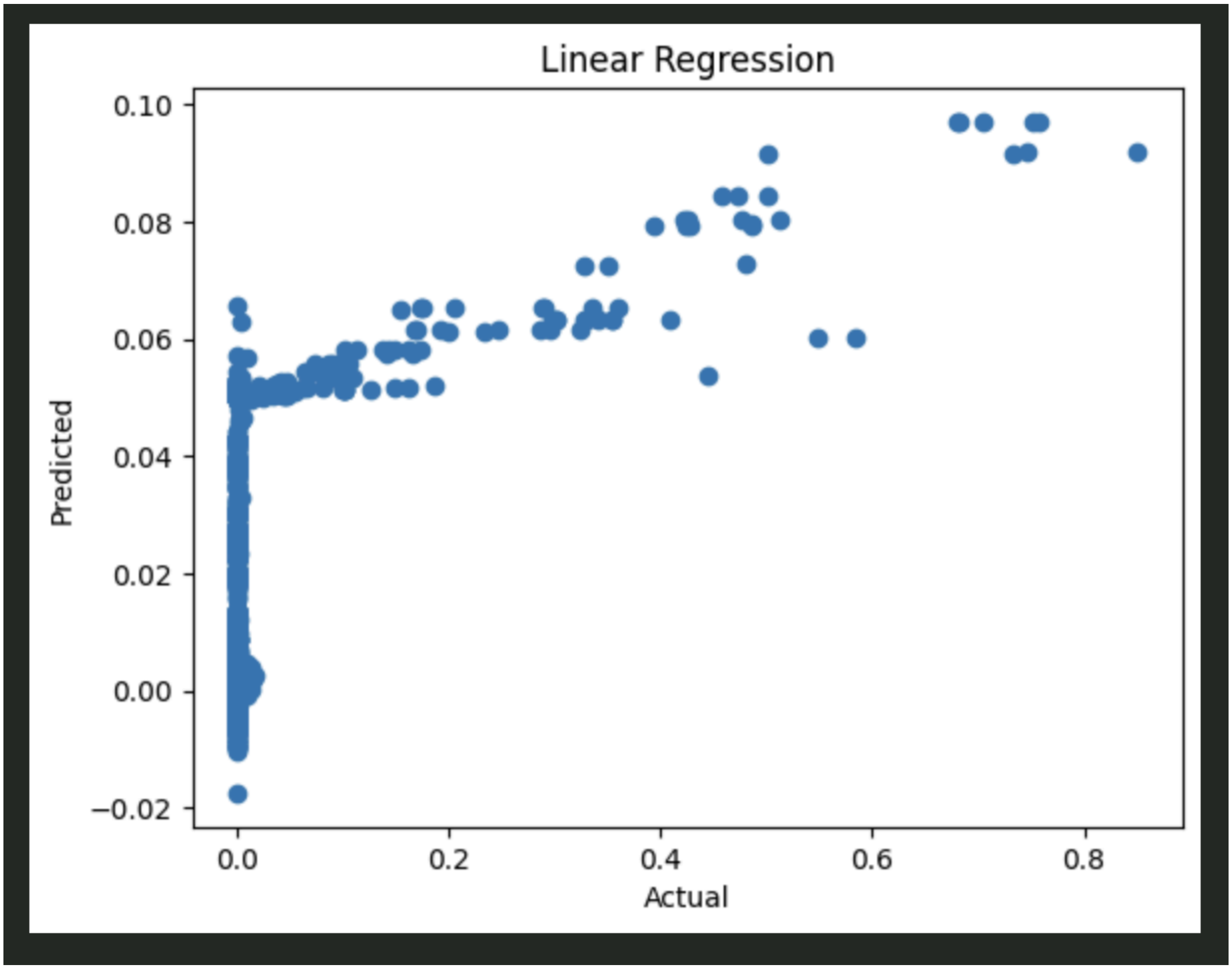
1. Splitting the Dataset
    a. Observation: The dataset was split into training and testing sets.
2. Training the Linear Regression Model
    a. Observation: A Linear Regression model was trained on the training data.
3. Making Predictions
    a. Observation: The trained model was used to predict Production on the test set.
4. Evaluating the Model
    a. Observation: The model's performance was evaluated using the R^2 score.



```
from sklearn.metrics import r2_score
r = r2_score(y_test,lr_predict)
print("R2 score : ",r)
```

```
R2 score :  0.1783199677700671
```

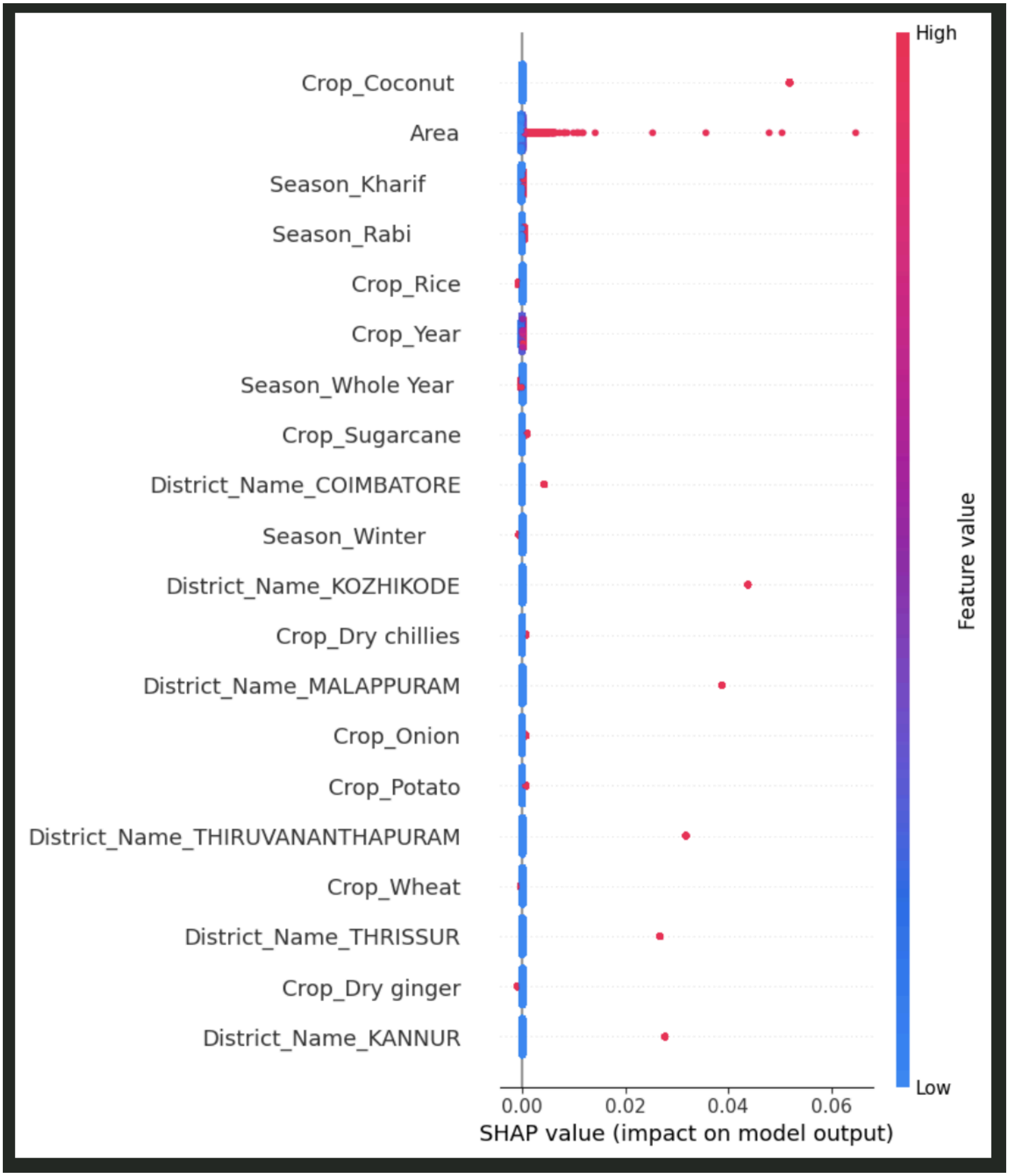
5. Visualizing Results
    a. Observation: A scatter plot was created to compare actual vs. predicted values.



6. Saving the Model
    a. Observation: The trained model was saved for future use.

Observations from SHAP Summary Plot

The SHAP (SHapley Additive exPlanations) summary plot provides insights into the impact of each feature on the model's predictions.



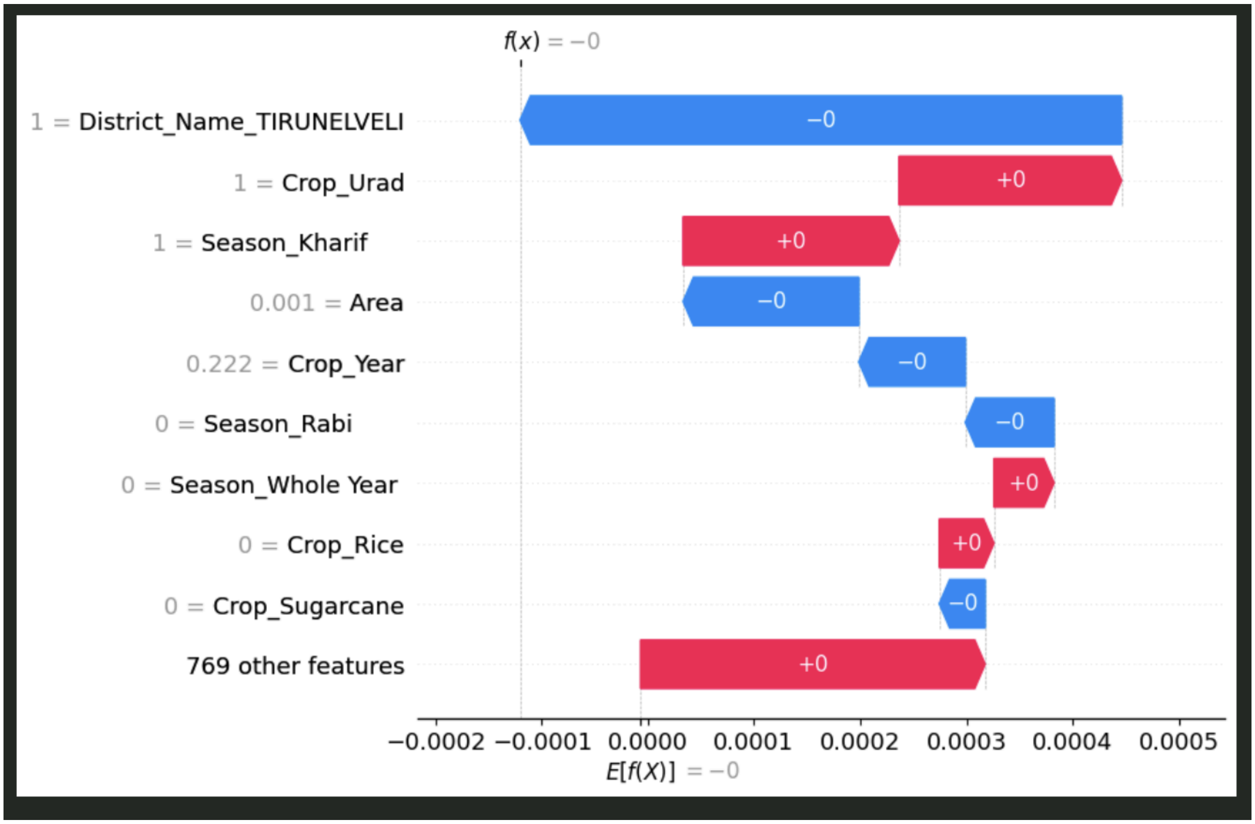
1. **Feature Importance -** The plot ranks features by their importance, with the most important features at the top.
    a. Observation: Features like Crop_Coconut, Area, Season_Kharif, and Crop_Rice appear to have the highest impact on the model's predictions.

2. **Direction of Impact -** The color of the points indicates the value of the feature (red for high values, blue for low values).
    a. Observation:
        i. **For Area:** High values (red) are associated with higher SHAP values, indicating that larger cultivation areas lead to higher predicted production.
        ii. **For Crop_Coconut:** High values (red) are associated with higher SHAP values, suggesting that coconut cultivation has a positive impact on production.
3. **Feature Distribution -** The spread of points for each feature shows the distribution of its impact.
    a. Observation:
        i. **Area** has a wide spread of SHAP values, indicating that its impact varies significantly across different instances.
        ii. **Season_Kharif** has a more concentrated impact, suggesting a more consistent influence on predictions.
    b. Explanation: Features with a wide spread have a variable impact depending on other factors, while those with a narrow spread have a more consistent effect.

The SHAP summary plot reveals that features like **Area and Crop_Year** are the most influential in predicting crop production.

Observations from SHAP Waterfall Plot

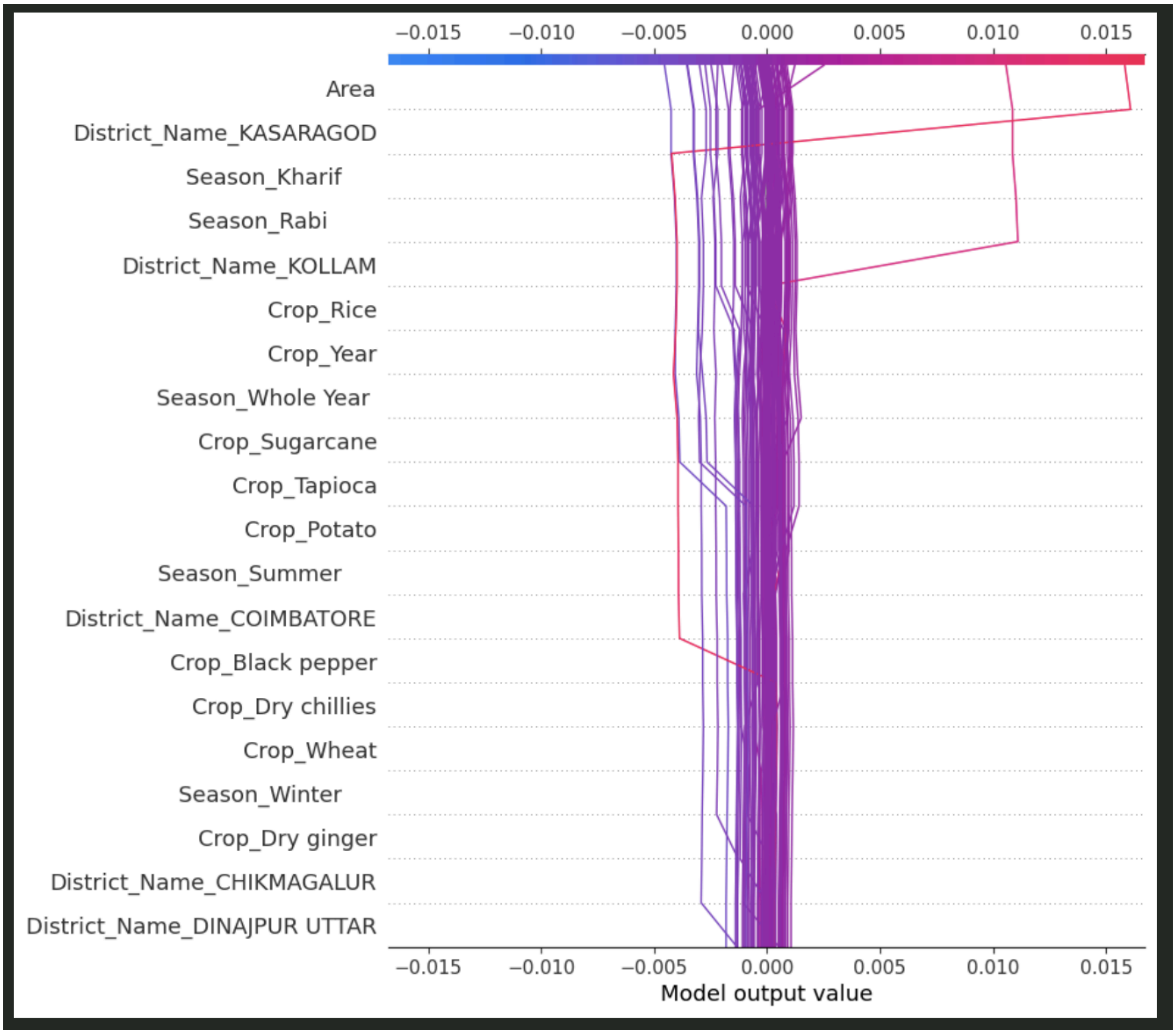
The SHAP waterfall plot provides a detailed breakdown of how each feature contributes to the model's prediction for a specific instance.



1. **Base Value -**
    a. Observation: The base value E[f(x)] is the model's average prediction over the training dataset.
    b. Explanation: This serves as the starting point for the prediction. In this case, the base value is approximately -0.0003.
2. **Feature Contributions**
    a. Observation: Each feature's contribution to the prediction is shown as a bar, with the length indicating the magnitude and the color indicating the direction (positive or negative).
    b. Key Contributions:
        i. District_Name_TIRUNELVELI: This feature has a significant positive contribution to the prediction.
        ii. Crop_Urad: This feature also has a notable positive impact.
        iii. Season_Kharif: This feature contributes positively but to a lesser extent.
        iv. Area: This feature has a minimal positive contribution.
        v. Crop_Year: This feature has a small positive contribution.
3. **Negligible Contributions**
    a. Observation: Features like **Season_Rabi, Season_Whole Year, Crop_Rice, and Crop_Sugarcane** have zero or near-zero contributions.

Observations from SHAP Decision Plot

The SHAP decision plot provides a visual representation of how the model's predictions are influenced by the features for multiple instances.



1. **Expected Value**
    a. Observation: The plot starts with the expected value E[f(x)], which is the average prediction over the training dataset.
    b. Explanation: This serves as the baseline for understanding how each feature shifts the prediction for individual instances.
2. **Feature Contributions**
    a. Observation: Each line in the plot represents an instance, and the features are listed along the y-axis. The x-axis shows the SHAP values, indicating the contribution of each feature to the prediction.
    b. Key Contributions:

        i. Area: This feature has varying impacts across different instances, with both positive and negative contributions.
        ii. District_Name_KASARAGOD: This feature shows significant positive contributions for some instances.
        iii. Season_Kharif: This feature generally has a positive impact on the predictions.
        iv. Crop_Rice: This feature has mixed contributions, with some instances showing positive and others showing negative impacts.

3. **Instance Variability**
    a. Observation: Different instances show varying patterns of feature contributions.
    b. Explanation: This variability indicates that the impact of features depends on the specific characteristics of each instance. For example, the impact of Crop_Rice may vary based on other features like Season or District_Name.