1.Preprocessing

a). Dataset is Loaded as pandas Dataframe b). Feature Selection :

- "hsi_id" column is dropped since it was row identifier.
- target variable is "vomitoxin_ppb".

c)Train- Test split:

- 80% of the data is used for training, and 20% is used for testing.

d)Missing Value:

- There was no missing values.

e) Outlier Handling

- Outliers were identified using IQR and then capped.

f) Dimensionality Reduction

- PCA was applied and 4 PC were generated which explained 96% variance.
- PC1 explained around 87% variance.

2. Model Selection

a)Linear Regression

- Evaluated using R^2 score and cross-validation.
- Cross- validation provided a negative R2 score.

b) Decision Tree Regression

- Achieved R^2 on training data: 1 and Test R^2: 0.1 suggesting overfitting and poor generalization.

3.Key Findings

- Output variable is higly skewd thats why linear regression is not working well.
- Applying other models like random forest or neural network and also tuning hyperperameter may improve R2 score.