Project Description:

The dataset used for predictive modeling was generated by the Wild Blueberry Pollination Simulation Model, which is an open-source, spatially-explicit computer simulation program, that enables exploration of how various factors, including plant spatial arrangement, outcrossing and self-pollination, bee species compositions and weather conditions, in isolation and combination, affect pollination efficiency and yield of the wild blueberry agroecosystem. The simulation model has been validated by the field observation and experimental data collected in Maine USA and Canadian Maritimes during the last 30 years and now is a useful tool for hypothesis testing and theory development for wild blueberry pollination research. This simulated data provides researchers who have actual data collected from field observation and those who want to experiment with the potential of machine learning algorithms response to real data and computer simulation modeling generated data as input for crop yield prediction models.

Problem statement:

The target feature is **yield** which is a continuous variable. The task is to classify this variable based on the other 17 features step-by-step by going through each day's task. The evaluation metrics will be the RMSE score.

Project Objectives:

1. EDA using matplotlib, pandas, and seaborn
2. Feature selection using `mutual_info_regressor`
3. Clustering to cluster types of bee columns
4. Standardizing input features
5. Baseline modeling using gradient-boosted trees:
6. Cross-validation using gradient boosted trees:
7. Model hyperparameters tuning using pipeline object with XGBRegressor:
8. Explainable AI using `shap`