**Dataset Preparation for Crop Type Classification**

This document describes the sequence of Jupyter notebooks used to clean, filter, and structure the dataset before calibrating the crop type classification model using XGBoost. The goal is to ensure a high-quality, temporally-consistent dataset ready for training and testing.

**1\_process1\_10week.ipynb**

This notebook performs the initial cleaning and aggregation of raw data:

* Load raw field-level data with weekly crop and index values.
* Filter based on coverage and NDVI quality thresholds.
* Compute summary statistics (mean, max, min, std) for vegetation and climatic indices.
* Export a cleaned DataFrame covering all 10 weeks per field-year.

**1\_process2\_10week.ipynb**

This notebook structures the cleaned dataset into a long format suitable for time series analysis:

* Load the output from the previous notebook.
* Flatten the weekly structure so each row represents a field-year-week combination.
* Add metadata such as FIELDID, Year, and Week for merging and identification.

**2\_filter\_dates.ipynb**

This notebook filters out invalid weekly entries that fall outside the typical growing season:

* Define valid time windows (e.g., sowing to harvest period) for each crop.
* Remove out-of-season or anomalous weeks.
* Ensure consistent weekly representation across samples.

**3\_reduce\_25week.ipynb**

This notebook produces the final machine learning dataset:

* Select relevant features (e.g., NDVI, EVI, precipitation, soil moisture).
* Drop incomplete records (with NaNs or irrelevant columns).
* Append label information (e.g., Crop\_num).
* Export final dataset for model training in 1\_SB25r\_Fit\_Classifier.ipynb.

**General Instructions**

**Environment Setup:**

* Required Python packages: pandas, numpy, matplotlib, seaborn.
* Use **Jupyter Notebook** for running and visualizing .ipynb notebooks interactively.

**Workflow:**

* Run 1\_process1\_10week.ipynb to clean and aggregate raw field-level data.
* Use 1\_process2\_10week.ipynb to convert the data into a structured weekly format.
* Run 2\_filter\_dates.ipynb to retain only valid in-season observations.
* Use 3\_reduce\_25week.ipynb to finalize the dataset for model calibration.

**Outputs:**

* Structured weekly crop dataset.
* Filtered and cleaned training dataset.
* Final labeled .csv file used in 1\_SB25r\_Fit\_Classifier.ipynb.