# Crop-Type Data Preparation — Simple Walkthrough

This document explains, in simple language, what each of the four main notebooks does. They are part of a pipeline that prepares data for crop-type prediction models. No coding knowledge is required to understand this summary.

## Overview

The notebooks clean, organize, and standardize satellite-based data (from Sentinel-2 and Sentinel-1) so that it can be used to train machine learning models that identify crops from space. The process transforms raw weekly observations into a clean, balanced, and consistent dataset with the same number of weeks (25) for each agricultural field.

The workflow follows this order:  
1. process1\_10week → Clean and prepare data from early seasons  
2. process2\_10week → Clean and prepare data from later seasons  
3. filter\_dates → Keep only relevant weeks for each crop and add simple features  
4. reduce\_25week → Standardize every sample to a fixed 25-week format

## 1. process1\_10week.ipynb

• Combines several years of satellite data into a single table.  
• Corrects crop labels that were mixed (e.g., 'Wheat + Maize'). It uses the week of the year to decide which label is correct.  
• Keeps only good-quality data (NDVI ≥ 0.2, meaning clear vegetation signal).  
• Keeps only fields that have at least 10 valid weeks of vegetation and radar data.  
• Balances the dataset so that very common crops do not dominate the model.  
→ Output: SB10r\_n0.2\_process1.csv

## 2. process2\_10week.ipynb

• Similar to the first notebook, but applied to later agricultural seasons.  
• Follows the same logic of cleaning, filtering, and balancing data.  
• Ensures each crop-year has a fair number of samples for model training.  
→ Output: SB10r\_n0.2\_process2.csv

## 3. filter\_dates.ipynb

• Merges the two cleaned datasets from steps 1 and 2.  
• Removes weeks that fall outside of each crop’s growing season:  
 – Wheat: keeps roughly weeks 17–47.  
 – Summer crops (Maize, Soy, Sunflower, etc.): keeps weeks within 22–40.  
 – Lucern: uses different weeks depending on irrigation.  
• Adds new features like week sine/cosine, growth rate, and NDVI/EVI ratios.  
• Helps visualize seasonal NDVI patterns for each crop.  
→ Output: SB10r\_n0.2\_process\_filt.csv

## 4. reduce\_25week.ipynb

• Makes sure every field’s time series has the same number of weeks (25).  
• Converts the time series data into a uniform structure, such as week\_01 to week\_25.  
• This fixed-length structure is required for most machine learning models.  
→ Final output: a uniform dataset ready for crop-type prediction.

## Summary

In short, these notebooks take raw, irregular, multi-year satellite observations and turn them into a structured, balanced, and ready-to-use dataset. This process ensures that the model can fairly compare crops and learn the typical seasonal patterns of each crop type.

## Key Terms

• NDVI / EVI / LAI – Vegetation indices that describe how green or dense the plants are.  
• SAR / Sentinel-1 – Radar data that can see through clouds.  
• Valid weeks – Weeks that fall inside the normal growing season for each crop.  
• Rebalancing – Reducing overrepresented crops so each class has a fair number of samples.