

2023 EY Open Science Data Challenge - Level 1

To develop a predictive model that can detect the presence, or non-presence of rice field at a given location based on satellite data. (Binary classification)

Dataset

Latitude and Longitude	Class of Land
(9.98941462845174, 105.4991893167462)	Rice
(9.97351661941364, 105.53643608077667)	Rice
(9.97533353473228, 105.52144652939856)	Rice
(9.9671574157984, 105.49373857079053)	Rice
(9.967611644628061, 105.50827389333898)	Rice
(9.982601196006842, 105.51690424110214)	Rice
(9.9830554248365, 105.53825299609522)	Rice
(9.968520102287382, 105.4946470284498)	Rice
(9.974425077072961, 105.50645697802042)	Rice
(9.96397781399078, 105.52371767354674)	Rice
(9.991685772600041, 105.51645001227251)	Rice
(9.985780797814462, 105.50872812216863)	Rice
(9.981238509517862, 105.50373160504259)	Rice
(9.971699704095002, 105.4923758843016)	Rice
(9.95307632207894, 105.51599578344286)	Rice
(9.97987582302888, 105.51554155461324)	Rice
(10.481616988271316, 104.91150805695796)	Non Rice
(10.481162759441657, 104.91150805695796)	Non Rice
(10.480708530611997, 104.91150805695796)	Non Rice
(10.480254301782336, 104.91150805695796)	Non Rice
(10.479800072952678, 104.91150805695796)	Non Rice
(10.479345844123017, 104.91150805695796)	Non Rice
(10.478891615293357, 104.91150805695796)	Non Rice



Satellite data for the 600 data points are sampled based on various conditions to provide comprehensive information for the model to classify the rice fields.

Satellite Data Collection

The satellite data of various wavelength bands and dates is accessed through Microsoft Planetary Computer Hub stac API with Python.

Range of Dates: Jan 2021 - Dec 2021

Data Sources

Sentinel-1 (radar data): VH, VV (can penetrate cloud (useful for target site which has high moisture and is often cloudy))

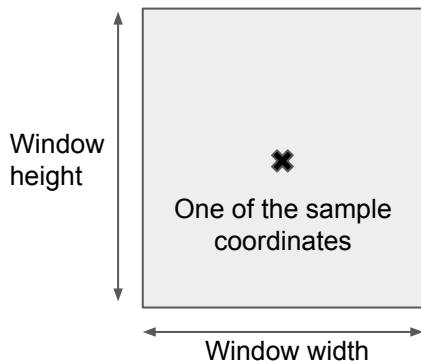
Sentinel-2 (optical data) data: B01, B02, B03, B04, B05, B06, B07, B08, B8A, B09, B11, B12, SCL, WVP, visual. (more informative visually but unable to penetrate cloud)

Collected Data

Satellite data collected is saved in multiple files with each of them containing satellite data of a particular source (s1 or s2) and date. e.g. LEVEL1_SENTINEL-2-L2A_2021-02-09 with satellite data collected from Sentinel 2 for 2021-02-09 along with their corresponding coordinates.is shown below.

Latitude and Longitude	lass of Lan	AOT	B02	B03	B04	B08	WVP	visual	AOT_w5	B02_w5	B03_w5	B04_w5	B08_w5	WVP_w5	visual_w5
(10.323727047081501, 105.2516346045924)	Rice	204	437	550	343	4212	4246	35	[[204. 204	[[452. 428	[[668. 639	[[385. 378	[[5128. 46	[[4560. 42	[[40. 39. 4
(10.322364360592521, 105.27843410554115)	Rice	204	1440	1544	1238	3450	4377	126	[[204. 204	[[693. 87	[[780. 94	[[617. 81	[[2426. 26	[[4121. 41	[[63. 83.
(10.321455902933202, 105.25254306225168)	Rice	204	431	594	326	4380	4750	34	[[204. 204	[[428. 433	[[511. 508	[[314. 301	[[3294. 32	[[4631. 46	[[32. 31. 3
(10.324181275911162, 105.25118037576274)	Rice	204	513	735	427	5756	4311	44	[[204. 204	[[492. 519	[[648. 658	[[381. 394	[[5424. 54	[[4275. 42	[[39. 41. 5
(10.324635504740822, 105.27389181724476)	Rice	204	3046	3276	3244	6496	4377	255	[[204. 204	[[2264. 30	[[2410. 32	[[2320. 28	[[5868. 62	[[4377. 43	[[236. 255
(10.323727047081501, 105.28070524968936)	Rice	204	385	467	326	2234	4093	34	[[204. 204	[[388. 386	[[463. 445	[[310. 292	[[2150. 20	[[4038. 40	[[32. 30. 3
(10.325089733570481, 105.23937042619212)	Rice	204	4796	4608	4078	7328	4377	255	[[204. 204	[[3554. 36	[[3200. 31	[[2546. 24	[[6480. 65	[[4377. 43	[[255. 246

Satellite data

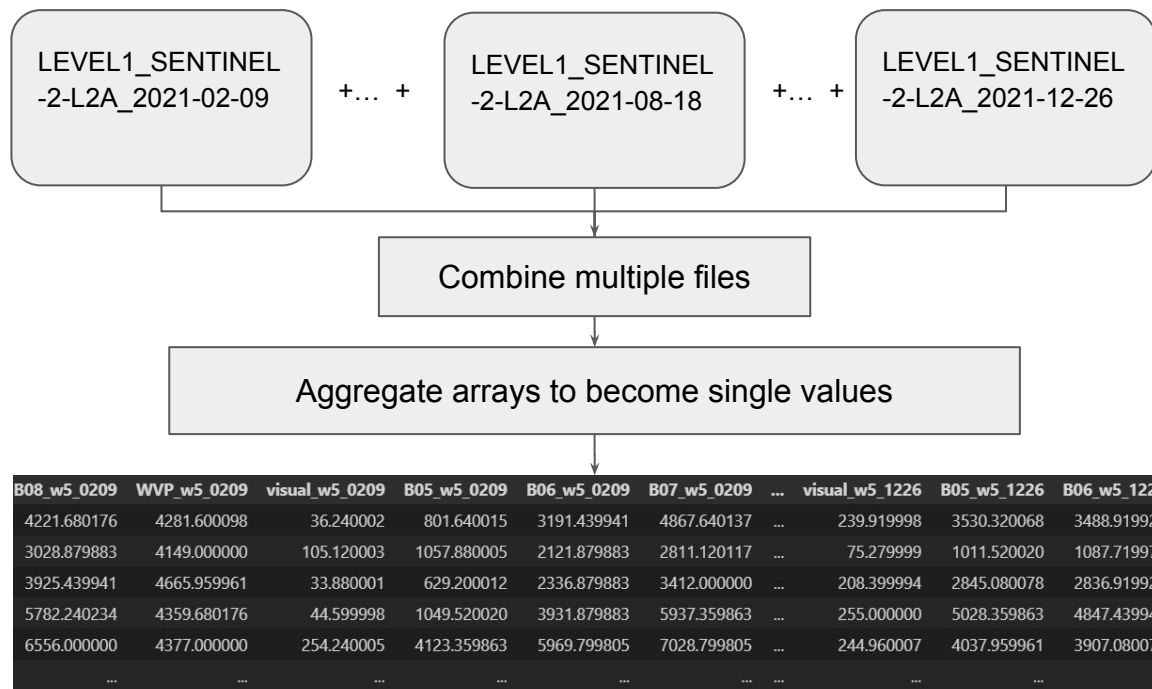


Each column indicates a column with specific wavelength band and window size, e.g. **B02_w5**:

- **B02**: wavelength band
- **W5**: window with width and height of 5 pixels

Each cell consists of an array of pixel values for its respective band and coordinate in the window.

Data Preprocessing



Sample of aggregated data.

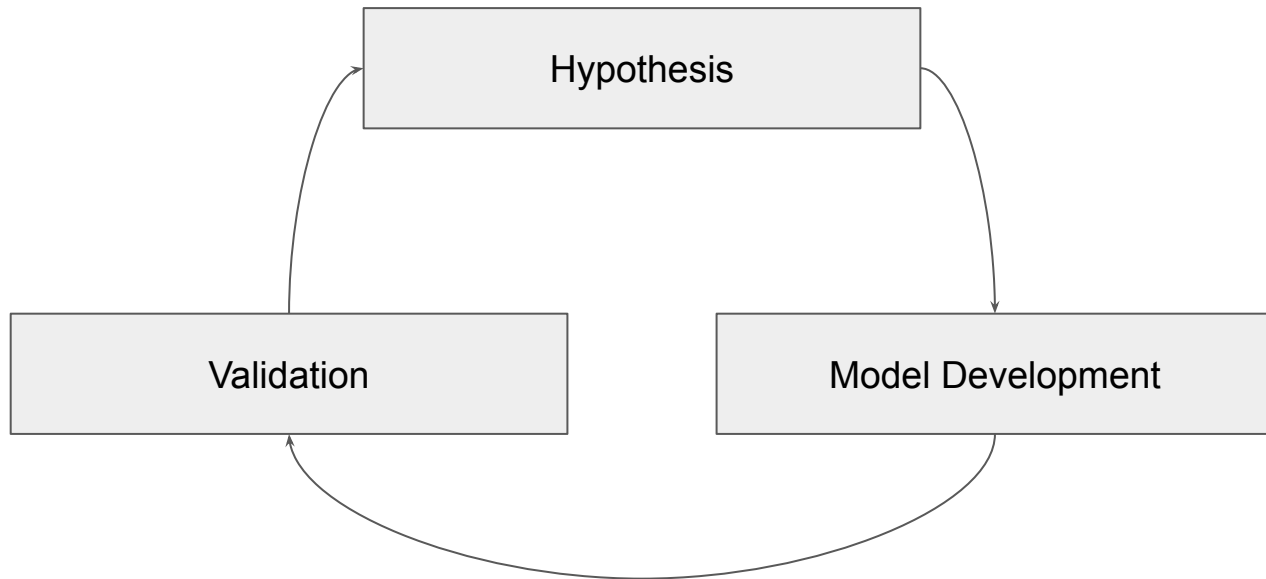
Multiple files are combined and aggregated to include information from multiple wavelength bands and dates as data for model development.

Each column contains data with specific wavelength, window size and date, e.g. **B02_w5_0209**:

- **B02**: wavelength band
- **W5**: original window width and height of 5 pixels which are aggregated into a single value with mean.
- **0209**: date of which satellite data is collected

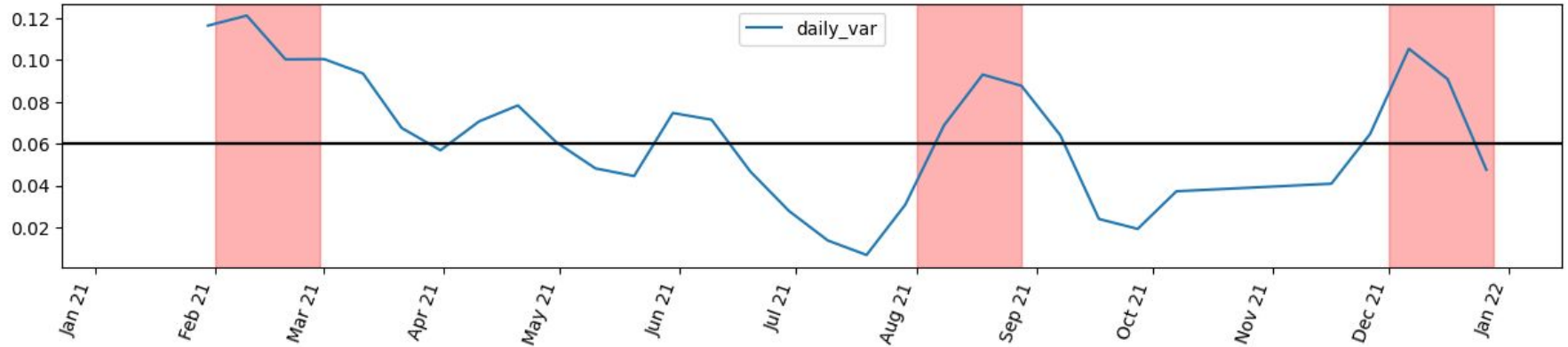
Input data is in a structured table format.

Model Development Experiments



Assumption 1: NDVI is the most important differentiator

In this case, we assume that the density of greenness of earth surface is the most important differentiator between different land covers. So, we selected data from the months with peaks in NDVI of rice fields compared with the average NDVI of all land covers, i.e. February, August and December. .



Evolution of Models + Test F1-scores

Based on the dataset queried based on the NDVI variance assumption, satellite data for February, August and December is used to develop classification models to predict the class of given coordinates

Iteration 1

test F1-score = 0.89

- Only Sentinel-2 data
- Random Forest
- window size of 5*5, aggregated by mean

Baseline model.

Iteration 2

test F1-score= 0.86

- all February, August, December data + NDVI
- window size of 5*5, aggregated by mean
- remove highly correlated features
- stack models

We assumed that stacking models + extracting NDVI might lead to improved performance increasing robustness. However, this approach reduced the F1-score when tested against unseen test data. These typical approaches are not the key to improve model performance.

Iteration 3

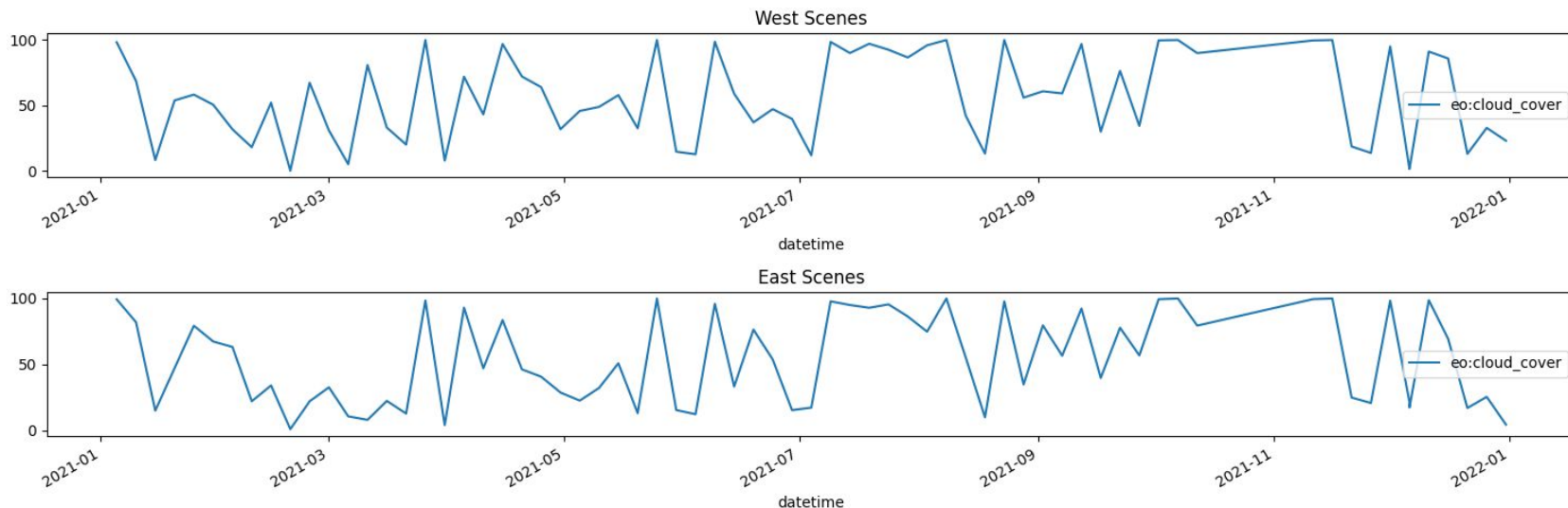
test F1-score= 0.95

- Sentinel-2: Feb, Aug, Dec raw data + NDVI
- Sentinel-1: Sentinel-1 raw data data + vh/vv
- drop highly correlated features
- normalization and stack models

Adding Sentinel-1 data to the configurations used in iteration 2 improved model performance significantly. New assumption: ability to remove the influence of cloud is the most important factor to increase model performance.

Assumption 2: Cloud Removal

As we observe that the inclusion of Sentinel-1 data in `iteration 3` significantly improves our prediction accuracy, we assume that **cloud removal is the most important factor in this classification task**. We conduct cloud analysis and find out the dates with the least amount of cloud hanging over the training sample coordinates throughout the year. Based on the dates, we extract their corresponding Sentinel-1 and Sentinel-2 data. We increased window size to 9*9 to reduce the impact of possible cloud covering the window.



Evolution of Models + Test F1-scores

Iteration 4

Test F1-score = 1.0 !!!

- take sentinel-1 and sentinel-2 data with <15% cloud coverage directly over our training data
- window size: 9*9, aggregated by mean
- normalization (MinMaxScaler)
- apply random forest classification

Based on the conditions established, we realise that we achieved a perfect F1-score of 1.0.

Iteration 4.1

Test F1-score = <1.0

- take sentinel-1 and sentinel-2 data with <15% cloud coverage directly over our training data
- window size: 9*9, aggregated by mean
- normalization (MinMaxScaler)
- remove highly correlated features
- apply random forest classification

Trying to simplify the classification model, removal of some highly correlated features results in reduction in test F1-score.

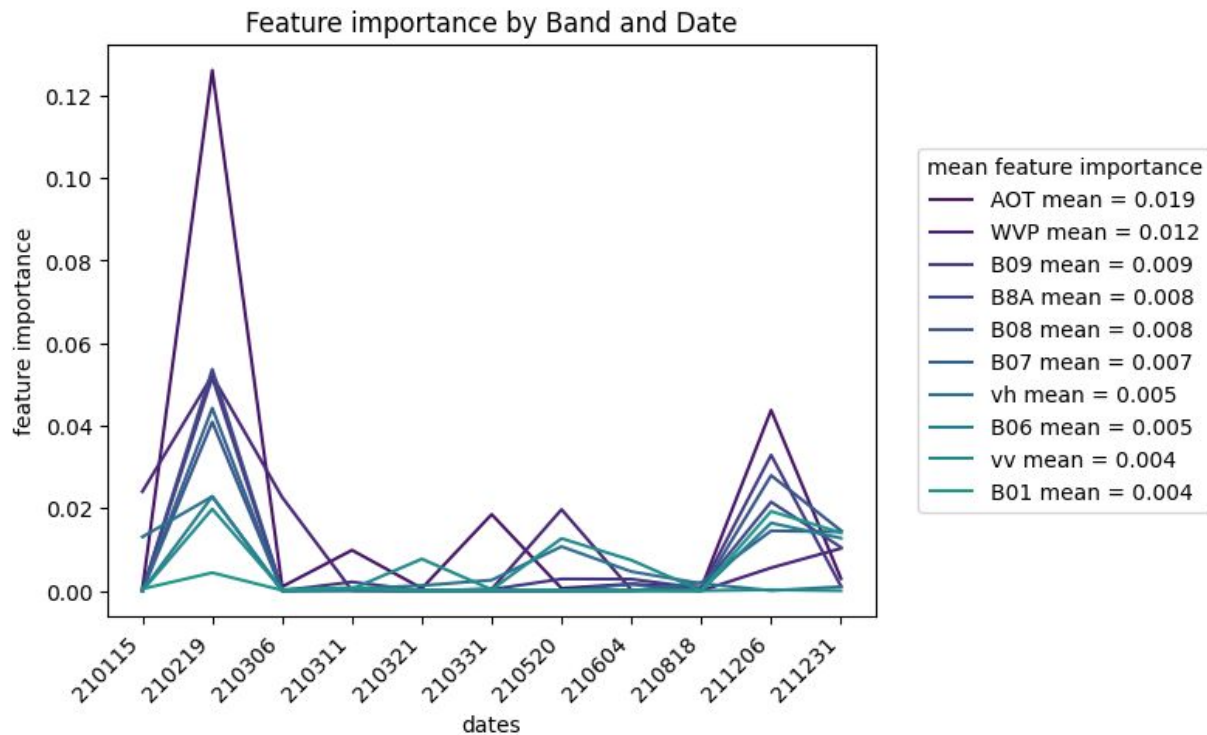
Iteration 4.2

Test F1-score = 1.0

- take sentinel-1 and sentinel-2 data with <15% cloud coverage directly over our training data
- window size: 9*9, aggregated by mean
- add NDVI and vh/vv features
- normalization (MinMaxScaler)
- remove highly correlated features
- apply random forest classification

By adding NDVI and vh/vv features to **iteration 4.1**, we achieve perfect F1-score of 1.0 again.

Highest Performing Features



Due to the large number of features and variation by time is involved, feature importance of the model is not easily interpretable.

In this case, we investigate the feature importance of each band for the dates chosen over the year. The plot shows the feature importances of the highest performing band over time.

Overall, 'AOT' and 'WVP' band have the highest average feature importance, but we notice that all bands have pretty small feature importance.