THE BRIDGE DIGITAL TALENT ACCELERATOR

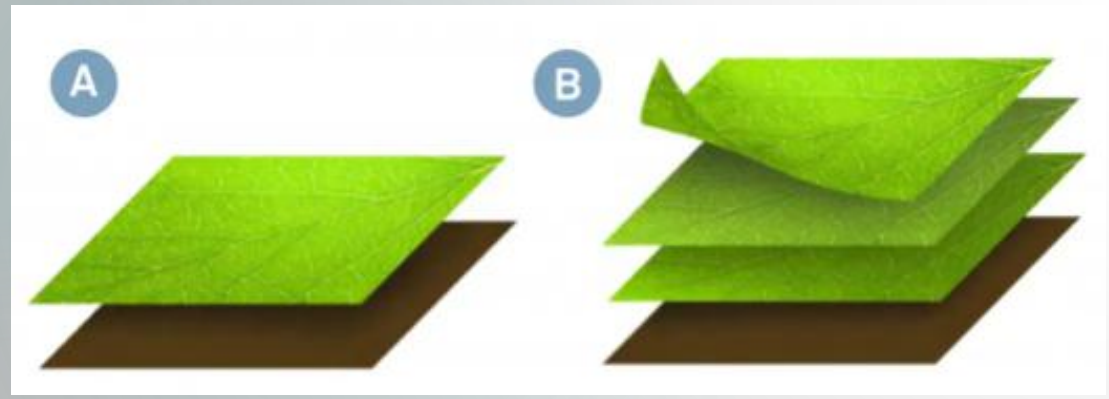
Desarrollo de un modelo de Machine Learning para la estimación del Indice de Area Foliar (LAI) a partir de imágenes de satélite

JOSE ESTEVEZ

Estimación de LAI y Machine Learning

Indice de área foliar
LAI [m^2/m^2]

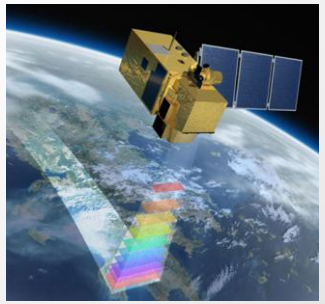
- Densidad de vegetación y capacidad fotosintética
- Crecimiento de cultivos y fenologia

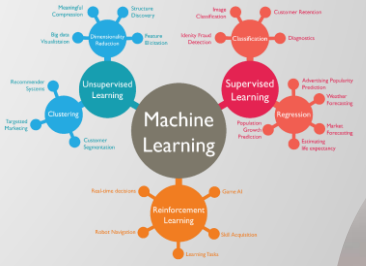


LAI = 1 LAI = 3

Métodos de medición:

- Directos
- Indirectos
  - Hemispherical Photography (DHP)
  - Transmittance
  - Reflectance





Recolección de datos

Datos insitu

DHP -> LAI

Proveedor:

GBOV
Ground-Based Observations for Validation
Copernicus Europe's eyes on Earth
Land Monitoring

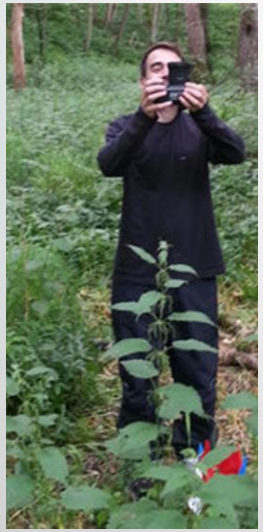
Fuente:

NSF | neon Operated by Battelle

29 sites

Landcovers:

- Evergreen Needleleaf
- Croplands
- Mixed Forest
- Deciduous Broadleaf
- Grasslands
- Evergreen Broadleaf
- Open Shrublands

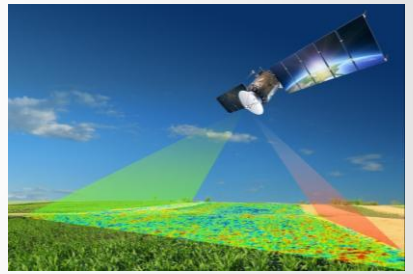


Datos de satélite

Reflectancia de superficie

Fuente:

Sentinel-2

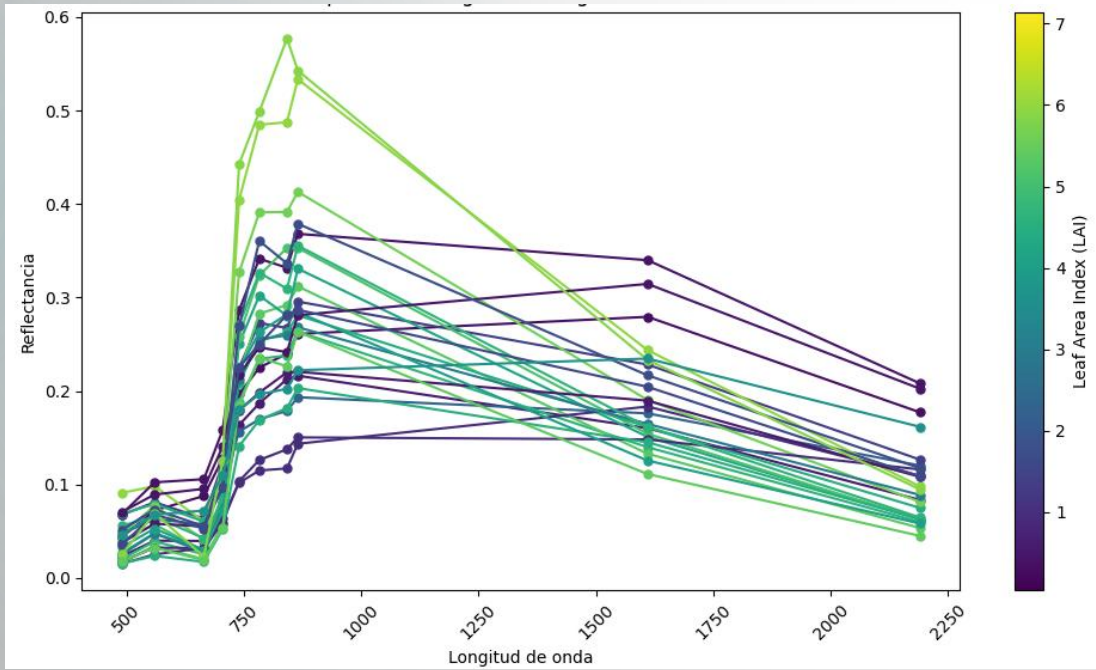


Proveedor:

Google Earth Engine

Unión de los datasets

	date_sat	date_insitu	longitude	latitude	B2	B3	B4	B5	B6	B7	B8	B8A	B11	B12	LAI_Warren
0	2019-06-07	2019-06-04	-72.171458	42.537834	0.0198	0.0532	0.0169	0.0982	0.3507	0.4087	0.4433	0.4232	0.1646	0.0673	4.030
1	2019-06-27	2019-07-02	-72.171458	42.537834	0.0236	0.0405	0.0193	0.0782	0.3244	0.4116	0.4295	0.4490	0.1786	0.0718	4.820
2	2019-07-12	2019-07-16	-72.171458	42.537834	0.0197	0.0203	0.0126	0.0247	0.0888	0.1154	0.0948	0.1156	0.0293	0.0117	3.871
3	2019-08-01	2019-07-30	-72.171458	42.537834	0.0176	0.0422	0.0188	0.0757	0.3112	0.3827	0.3980	0.3963	0.1670	0.0656	3.790
4	2019-08-26	2019-08-27	-72.171458	42.537834	0.0204	0.0405	0.0182	0.0711	0.2746	0.3450	0.3710	0.3663	0.1580	0.0626	4.281



Spectral bands		nm	m
~~B1~~	~~Coastal aerosol~~	~~443~~	~~60~~
B2	**Blue**	**490**	**10**
B3	**Green**	**560**	**10**
B4	**Red**	**665**	**10**
B5	**Vegetation red edge**	**705**	**20**
B6	**Vegetation red edge**	**740**	**20**
B7	**Vegetation red edge**	**783**	**20**
B8	**NIR**	**842**	**10**
B8a	**Narrow NIR**	**865**	**20**
~~B9~~	~~Water vapour~~	~~945~~	~~60~~
~~B10~~	~~SWIR-Cirrus~~	~~1375~~	~~60~~
B11	**SWIR**	**1610**	**20**
B12	**SWIR**	**2190**	**20**

Métodos

Método empírico

Datos insitu (LAI)

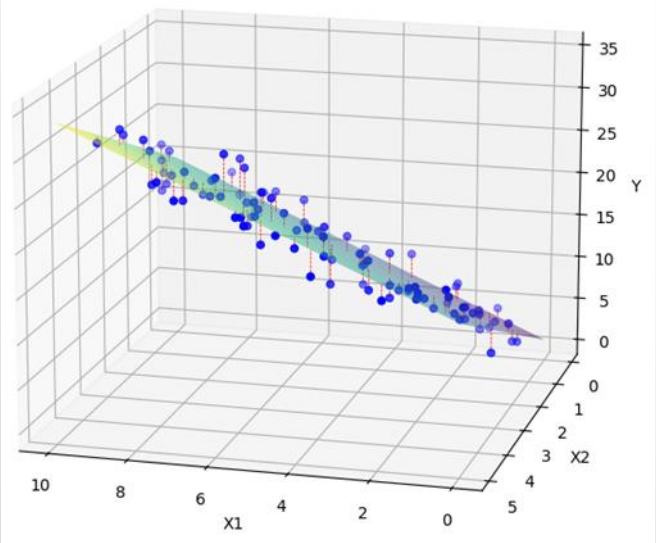
+

Datos de satélite (B2,B3,B4,B5,B6,B7,B8, B8A,B11,B12)

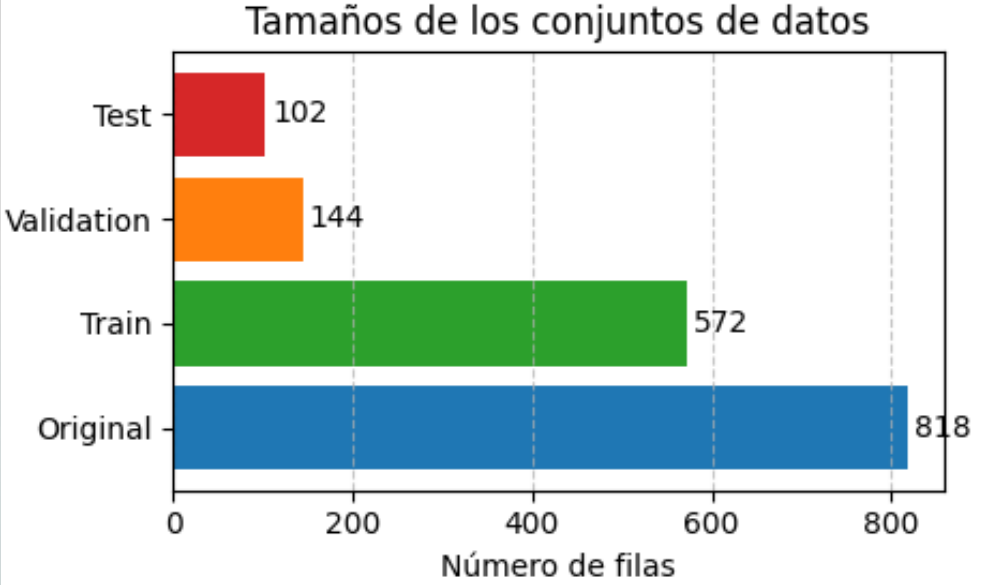
→ **Machine Learning (GPR)** →

LAI pred

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

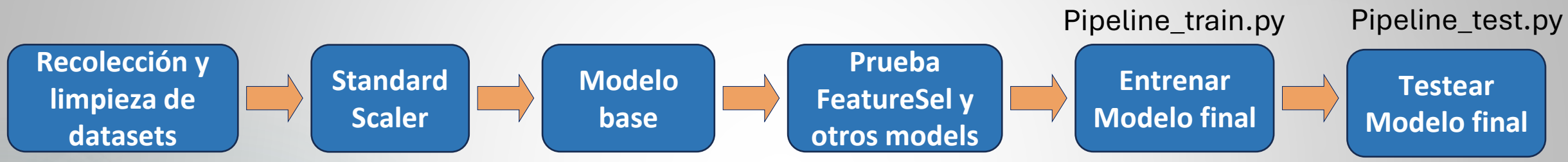


División del dataset



Test set -> 2023
Train set y **Val set** -> 2018 a 2022
(estratificación por tipo de cubierta)

Construcción del modelo de ML



Métricas usadas:

- RMSE
- **MAPE**
- UAR (Uncertainty Agreement Ratio)
- R^2

Feature selection y model testing

Feature reduction tecniques:

Visual, SelectedFromModel, RFE, SFS, Hard voting, Full_num_features

Machine Learning models:

Regresion Lineal, Ridge, Lasso, ElasticNet, SVR, RandomForest, XGBoost, LightGBM, CatBoost

6 * 9 = 54 modelos

Results

Top 4 Features mas predictivas
(hard voting)

- B7 (red edge)
- B8A (NIR)
- B11 (SWIR)
- B12 (SWIR)

Cross Validation

1	CatBoost-SFS: 0.57
2	**SVR-SFS: 0.57**
3	LightGBM-SFS: 0.57
4	SVR-Full_num_features: 0.58
5	SVR-SelectedFromModel: 0.59
6	SVR-RFE: 0.59
7	SVR-Hard voting: 0.59
8	Random Forest-SFS: 0.59
9	SVR-Visual: 0.60
10	Random Forest-SelectedFromModel: 0.61
31	**Regresion Lineal-Full_num_features: 0.75**

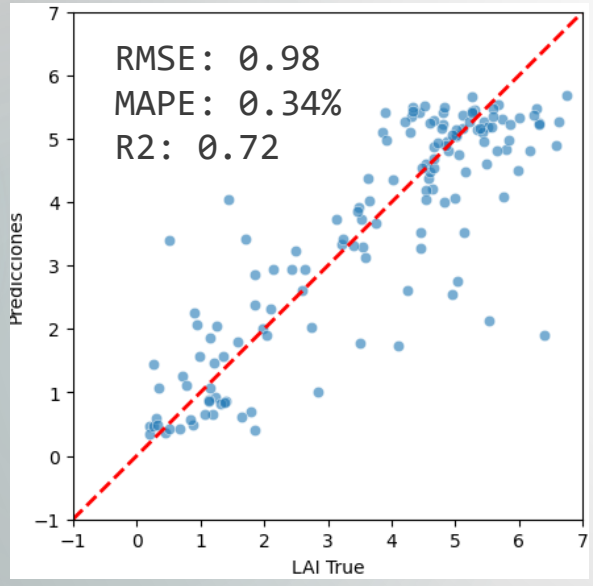
Preselección de modelos

Contra val_set

SVR-SFS

(6 Bandas)

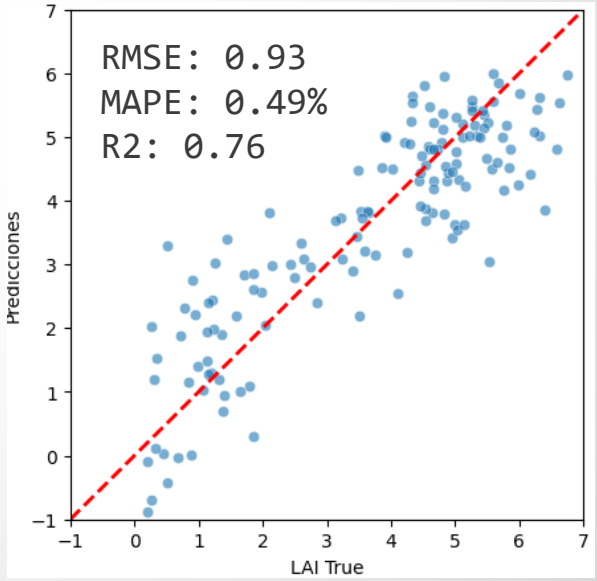
[B6, B7, B8, B8A, B11, B12]



Linear Regression

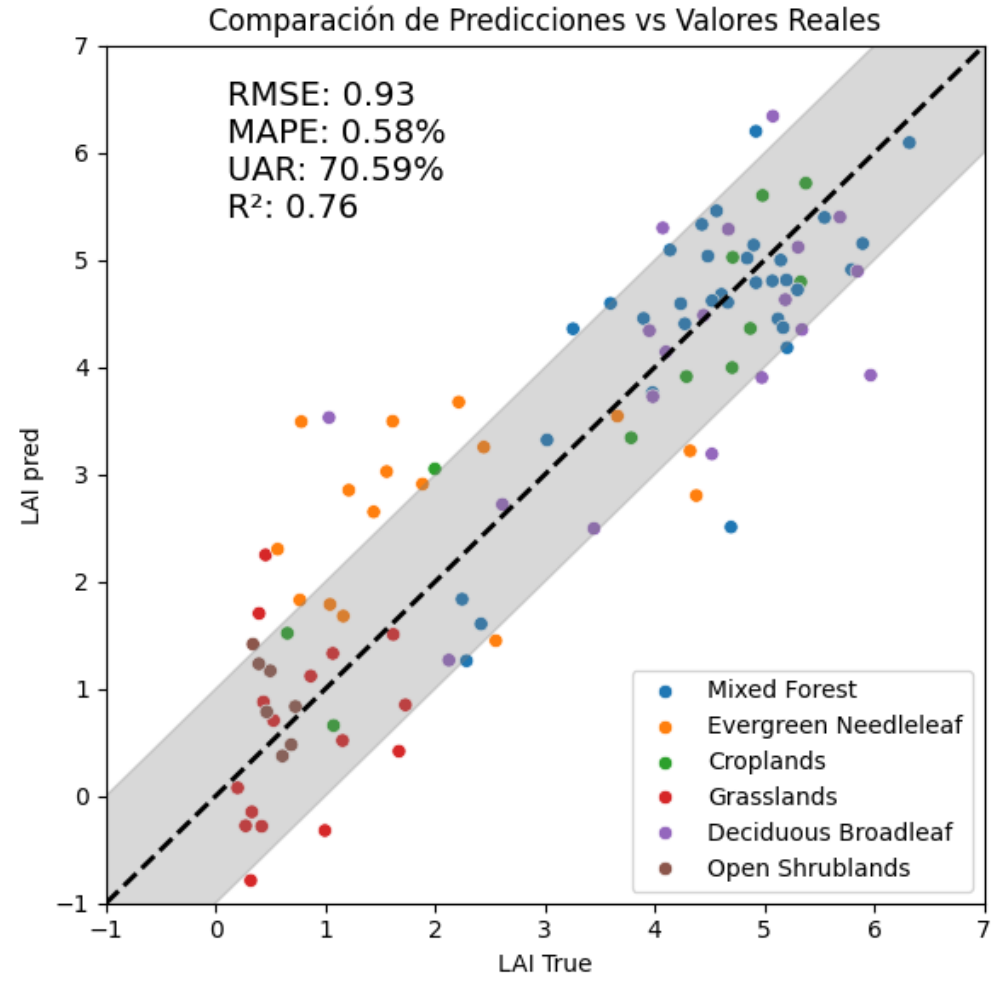
(10 bandas)

[B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12]



Modelo base

Modelo final

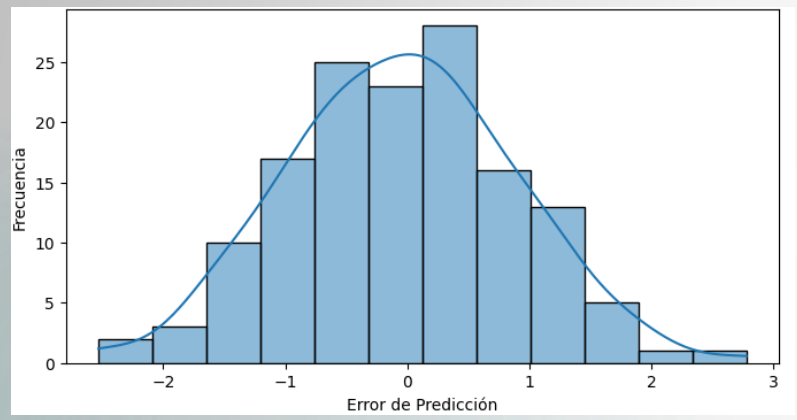


Ecuación de la regresión lineal:

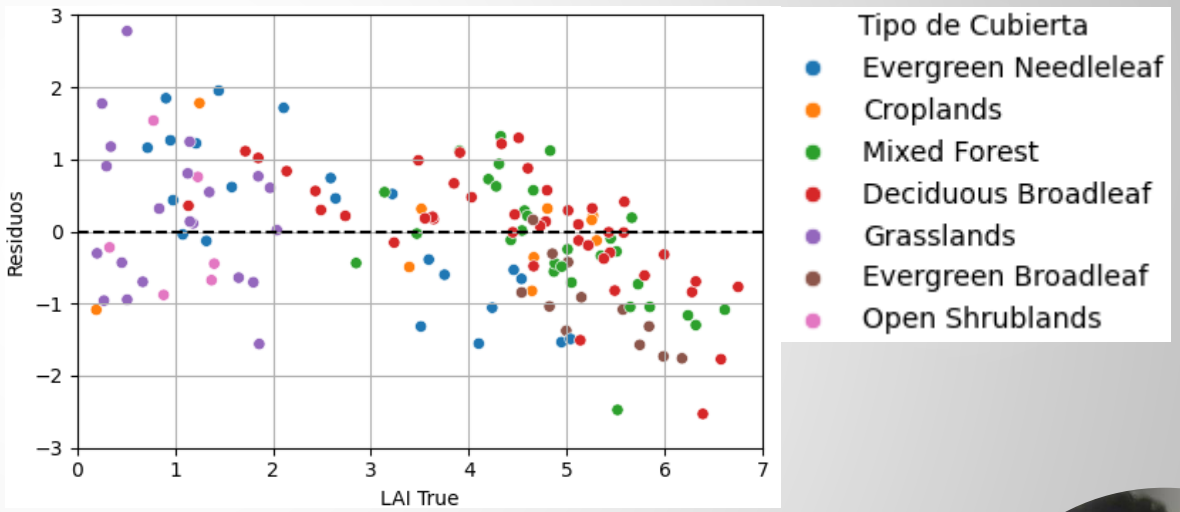
LAI_pred = 3.456 +
(0.5593 * B2) +
(-1.0306 * B3) +
(0.7463 * B4) +
(-0.3338 * B5) +
(1.4298 * B6) +
(-0.2352 * B7) +
(-0.0757 * B8) +
(0.5505 * B8A) +
(-1.7796 * B11) +
(-0.0398 * B12)

Análisis de errores del modelo

Distribución de residuos

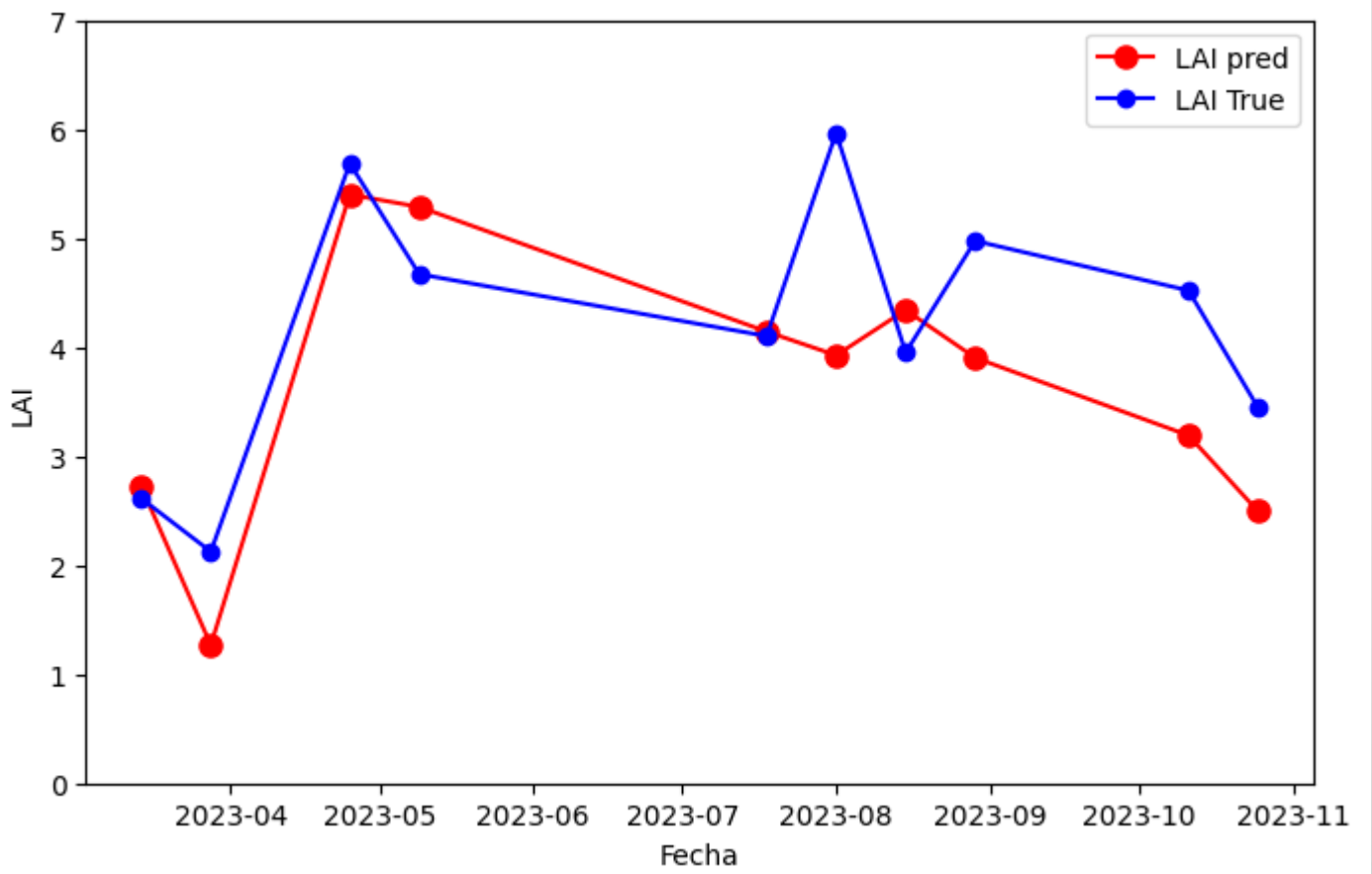


Análisis de residuos por cubierta



Aplicaciones: Fenología de la vegetación

Serie temporal de LAI

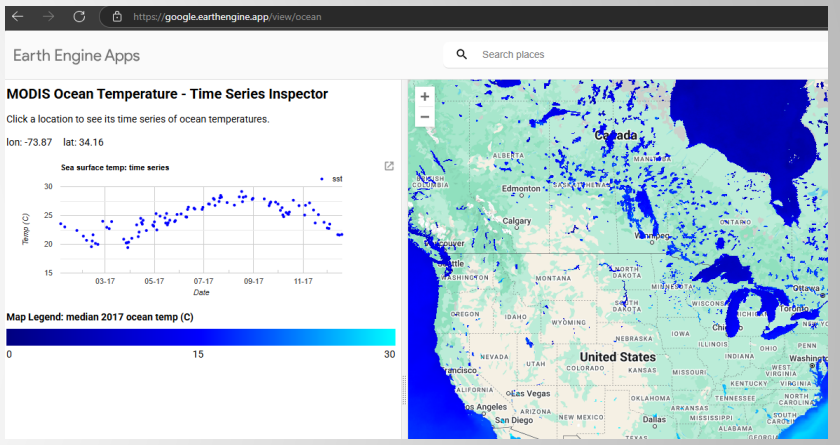


Conclusiones

- Modelo viable y mejorable para estimar LAI a gran escala espacial y temporal.
- La regresión lineal mostró desempeño un similar que modelos mucho más complejos.
- Las bandas del infrarrojo son más relevantes.
- La reducción de features no siempre mejora el modelo.
- Dificultad para predecir valores en los extremos.
- Los outliers disparan el RMSE.
- Unas cubiertas se estiman mejor que otras.
- Los errores se reducen con mejores datos de entrenamiento.

Acciones de mejora

- Conseguir más datos.
- Probar en otras zonas y a diferentes latitudes.
- Introducir ruido para generar nuevas muestras.
- Filtrar outliers.
- Reducir el umbral temporal.
- Forzar predicciones negativas a cero.
- Probar con PCA y GPR.
- Probar con reflectancias TOA.
- Implementar una web app (ej. Google Earth Engine Apps o Streamlit)

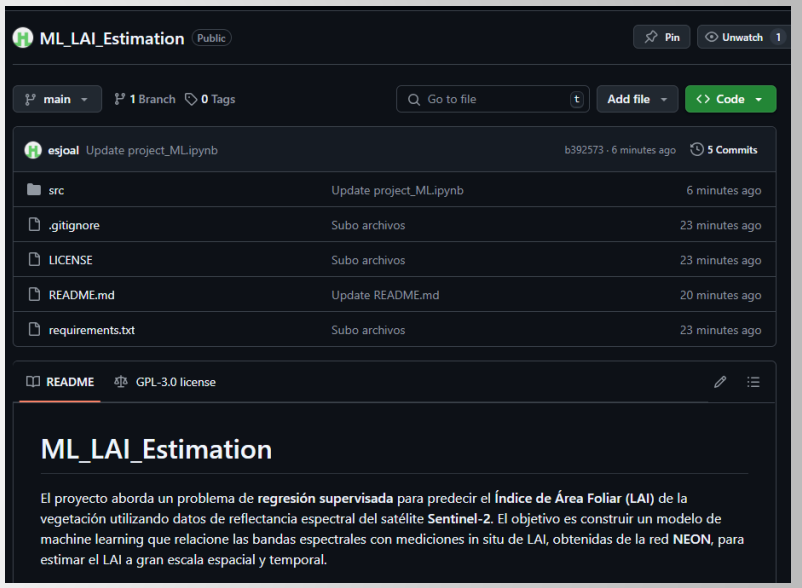


Código fuente

Estructura:

- Scripts para la recolección de datasets.
- Funciones para la limpieza, preprocesado y unión de los datasets.
- Pipelines para entrenar y testear el modelo.

https://github.com/esjoal/ML_LAI_Estimation/tree/main



¡Gracias!