Readme.

Reference to the associated papers

The code we describe in this document has been used to create the dataset *"A Multisource Grapevine Phenology Dataset for Smart Farming and AI Modeling"*. This dataset is available in [1], it is presented in [2] and has been used to create the grapevine phenology prediction models discussed in [3].

- [1] Francisco José Lacueva-Pérez et al. "*A multisource grapevine phenology dataset for smart farming and AI modeling*". Dataset on Zenodo. 2026. doi: 10.5281/zenodo.17930722. url: https://doi.org/10.5281/zenodo.17930723.
- [2] Francisco José Lacueva-Pérez et al. « A multisource grapevine phenology dataset for smart farming and AI modeling". UNDER REVISION.
- [3] *Francisco José Lacueva-Pérez et al. "Developing machine learning models from multisourced real-world datasets to enhance smart-farming practices". In: Computers and Electronics in Agriculture 231 (Apr. 2025), p. 110018. issn: 0168-1699.doi:10.1016/j.compag.2025.110018.url: http://dx.doi.org/10.1016/j.compag.2025.110018*.
  This paper describes the design, development and evaluation of a set of Machine Learning models that predict the phenology of grapevines. The models were trained on multisourced data coming from 9 different data sources with different temporal and spatial resolutions. The authors evaluated and compared different machine learning algorithms to predict 9 different phenological stages of grapevines. The models that performed best also included data derived from Sentinel-2 images, which suggests that multispectral satellite images could be used to monitor and predict woody plant phenology. A key contribution of our proposal is the combination of multiple data sources and a fine-grained prediction aimed at distinguishing among 9 phenological states.

Acknowledgements

This research was partially funded by: Agridatavalue project, co-financed by the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101086461. It is also part of the project PID2020-113037RB-I00 (NEAT-AMBIENCE project), funded by MICIU/AEI/10.13039/501100011033. Besides the NEAT-AMBIENCE project, we also thank the support of the Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón (Department of Science, University and Knowledge Society of the Government of Aragon) to the following research groups: COSMOS with reference T64_23R and IODIDE with reference T17_20R.

Introduction

This document outlines the specific modules employed to construct the dataset used for training and evaluating the machine learning (ML) models presented and compared in the

paper. Figure 1 offers a comprehensive overview of the workflow, encompassing data acquisition from diverse sources, subsequent transformation into intermediate datasets, their integration into a unified dataset for ML model training, and the subsequent generation of predictions based on the selected optimal model. While the workflow is depicted as linear, for simplicity, the actual process was iterative, involving continuous refinement of interim datasets, ML models, and predictions. To guide our work, we adhered to the CRISP-DM methodology.

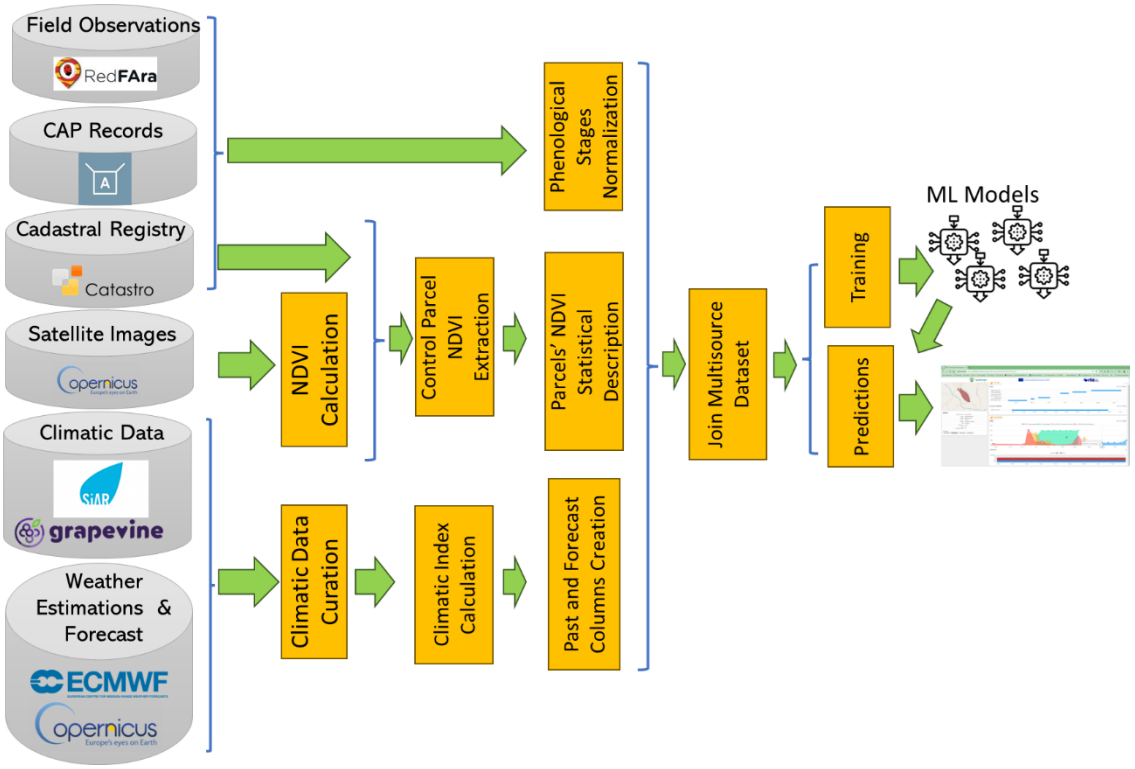

Figure 1 Workflow implemented to transform data used to create ML models compared in [1].

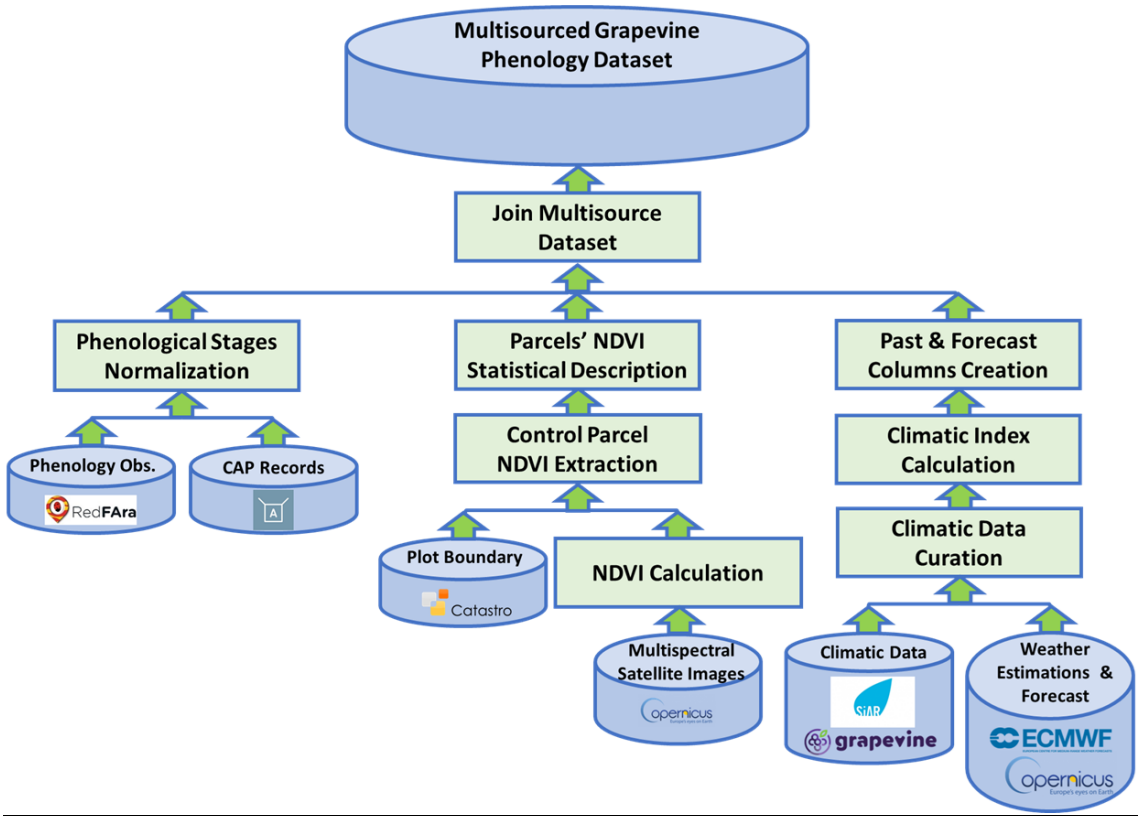

Figure 2. Subworflow for creating the dataset presented in [2].

Data Sources Considered

The following sections provide a brief overview of the data sources used in this study. All data sources are geographically linked to three Protected Designations of Origin (PDO) in Aragón (Spain): Calatayud, Campo de Borja, and Cariñena. These PDOs collectively encompass approximately 8,500 km^2 and are spatially defined by the WGS84 coordinates (41.98107, -2.177578) and (41.166320, -0.922575). Figure 1 shows the data sources grouped by the kind of data they provide. The description of the datasets is performed following the same classification:

- **Field Observations**. Phenological data is collected from the Ref FARA network, managed by agronomists at the Aragón Regional Government's Vegetal Health and Certification Center. Technicians visit selected parcels on a weekly basis to gather this information. Data are identified by the CAP (Common Agricultural Policy) code of each of the plotted parcel and the plotting date.
- **CAP Records**. Grapevine variety information is retrieved from publicly available Common Agricultural Policy (CAP) data for the plotted parcels. These data are accessed through the Aragón Open Data Portal. As expected, each parcel is uniquely identified by its CAP code and the date the data was recorded.
- **Spanish Cadastral Registry**. It provides geographical information of the parcel, which is identified by its cadastral code. There is a biunivocal transformation between the cadastral code and the CAP code of the parcels.
- **Satellite Images**. The multispectral images are obtained by the Copernicus Sentinel 2 satellite constellation and accessed through the Copernicus Data

Space. For creating the models, we downloaded and processed images covering the area of interest, which correspond to the tiles T31TBG, T30TYM, T30TXM, T30TXL, T30TWM, and T30TWL.

- **Climatic Data**. Real climatic variables are obtained from a couple of climatic station networks, which have stations deployed in or near the 3 POD: Grapevine and SIAR Networks.
- **Weather Estimations and Forecast**. To replace missing data and support future predictions, we used estimations and the weather forecast provided by the ECMWF ERA5 models. We used the Open Meteo API to have an easier access to the data than the one available with the ERA5 models results.

Source Code for Data Retrieval

Several modules are involved in gathering data from the different data sources:

- src/main/python/sarga/aemet2pg.py loads AEMET data (access restricted by IP).
- src/main/python/sarga/descsiar.py loads data from the SIAR climatic station network.
- src/main/python/sarga/gv2pg2.py updates data from the Grapevine climatic station network.
- src/main/python/ForecastAPI/OpenMeteoRecoverPrevoiusDataForCompletingDatabase.py downloads climatic data estimations to replace missing data.

Source Code for NDVI Calculation

The functionality to compute the NVDI is included in the following Python script:

- src/main/python/copernicus_new/main.py downloads Copernicus Sentinel 2 multispectral images, calculates the NDVI, prunes parcel areas and calculates the descriptive values of the NDVI (max, min, average, etc.).

Source Code for Climatic Data Curation and Climatic Indexes Calculation

The following Python script contains code for data curation and computation of climatic indexes:

- src/main/python/climaticDataSiarWinkler/TransferSiarDataNew.py calculates all the possible combinations of the GDD, Winkler, Richardson and UTAH indexes, as well as other derived variables which are included and described in the dataset provided together with this code. For calculating the indexes, the beginning of the season for each of the stations is also calculated. Phenological data are normalized using the mapping table included as part of the provided dataset.

Source Code to Join the Multisource Dataset

To combine and integrate the different input data sets, we use the following code:

- src/main/python/CreateFinalSheet/Tejedor_de_Parras_Structured.py merges the data prepared from each data source based on temporal and location data. The resulting dataset is the one provided together with this code.

Source Code for Training the Models

The folder src/main/python/nb_models contains the Python notebooks which are used to train the models based on the data included in the generated combined dataset that we share together with this code. For each algorithm, we create a notebook for training and validating the model and another to test the best performing model:

- Fenologia_models_all_data-GB-Test-parcels.ipynb
- Fenologia_models_all_data-GB-parcels.ipynb
- Fenologia_models_all_data-RF-parcels.ipynb
- Fenologia_models_all_data-RF-Test-parcels.ipynb
- Fenologia_models_all_data-NN-Test-parcels.ipynb
- Fenologia_models_all_data-NN-parcels.ipynb