Project Proposal - Canal Irrigation Prediction for the Ebro Basin, Spain

Course: Machine Learning

September 30, 2025

Emman Abrar Ali [23K-0051]

Areeba Hasnain [23K-0059]

Ibrahim Johar [23K-0074]

Amna Asim Khan [23K-0859]

Motivation

In semi-arid regions such as the Ebro Basin, water represents a scarce resource. Inefficient irrigation practices contribute to water waste, reduced crop yields, and economic losses for agricultural stakeholders. The objective involves transitioning from static, calendar-based irrigation approaches to dynamic, data-driven methodologies

Objectives

- Collect and merge datasets on weather, soil moisture, crop irrigation, and canal water availability in Spain.
- Build an ML model (classification: irrigate vs. don't irrigate, or regression: days until next irrigation).
- Evaluate performance using cross-validation and feature importance.
- Provide a prototype decision-support tool for irrigation scheduling in the Ebro Basin.

Data Sources (Spain – Open Access)

We will integrate the following datasets:

Weather & Agroclimate

SiAR – Agroclimatic Information System for Irrigation (Spain)

Provides weather data (rainfall, temperature, humidity, radiation, wind) from irrigated regions.

🔗 SiAR Portal – Datos abiertos del Gobierno de España | datos.gob.es

Soil Moisture

Copernicus Land Monitoring – Soil Moisture

High-resolution (5 km and finer) soil moisture maps for Europe, including Spain.

🔗 Copernicus Soil Moisture - https://climate.esa.int/en/projects/soil-moisture/

ESA CCI Soil Moisture (Global)

Long-term climate-quality soil moisture records can be subset for Spain.

🔗 ESA CCI Soil Moisture - https://climate.esa.int/en/projects/soil-moisture/

Irrigation & Canal Data

Ebro Basin Irrigation Water Dataset (1 km resolution, 2000–2020)

Satellite-derived irrigation water estimates over the Ebro River Basin. Includes irrigation demand and usage.

🔗 ESSD – Ebro Irrigation Dataset -https://essd.copernicus.org/articles/15/1555/2023/

Crop Irrigated Area

ECIRA – European Crop-Specific Irrigated Area Dataset (2010–2020, 1 km resolution)

Provides annual irrigated area by crop type in Spain.

🔗 Nature – ECIRA Dataset - https://www.nature.com/articles/s41597-025-05628-y

Methodology

I. Data Integration:

- Align weather, soil moisture, irrigation, and crop area data by time (daily/weekly) and space (1 km grid in Ebro Basin).

II. Feature Engineering:

- Weather: rainfall, evapotranspiration, temperature, solar radiation.
- Soil moisture: topsoil & root-zone levels.
- Canal/Irrigation: irrigation water estimates, canal availability.
- Crop type: from ECIRA dataset.

III. Modeling:

- Classification model: Predict “Irrigate today?” (Yes/No).
- Regression model: Predict “Days until next irrigation”.
- Algorithms: Random Forest, Gradient Boosted Trees (XGBoost), LSTM for time series.

IV. Evaluation:

- Metrics: Accuracy, Precision/Recall (for classification); RMSE, MAE (for regression).
- Compare with a rule-based baseline (FAO crop water requirement formula).

Expected Outcomes

- A machine learning model that suggests the optimal irrigation day for fields in the Ebro Basin.
- Demonstration of how open-access European datasets can be leveraged for water management.
- Potential for scaling the approach to other semi-arid agricultural regions.

Tools & Frameworks

- Language: Python
- Frameworks: Scikit-learn, XGBoost, PyTorch (if deep learning).
- Data Handling: Pandas, GeoPandas, NetCDF4, Rasterio.
- Visualization: Matplotlib, Plotly.

High-Level Approach:

Data Collection: Multi-source data including satellite imagery, weather records, and soil characteristics will be gathered for the Ebro Basin region.

Feature Engineering: These datasets will be combined to create engineered features such as soil moisture deficit, evapotranspiration rates, and crop water stress indices.

Model Training: A machine learning model will be trained to predict optimal irrigation timing. The "optimal day" may be defined as the period preceding significant plant water stress, derived from soil moisture thresholds or domain expertise.

Deployment & Visualization: An interactive dashboard will be developed enabling users to select individual fields and view irrigation recommendations for the upcoming week.

Detailed Technical Framework

I. Programming Language: Python

Python serves as the dominant language in data science due to its accessibility and extensive ecosystem of specialized libraries.

II. Core Machine Learning Frameworks

Scikit-learn: This library provides fundamental tools for classical machine learning and data analysis. It is ideal for establishing baseline models using algorithms such as:

Random Forest: Effective for tabular data, capable of handling non-linear relationships, and providing feature importance metrics

Gradient Boosting Machines: Often demonstrates superior accuracy compared to Random Forests

Support Vector Machines: Suitable for regression tasks
Primary use cases include predicting continuous values such as "days until next irrigation" or binary classifications such as "irrigate within 48 hours."

XGBoost: As an optimized implementation of Gradient Boosting, this framework frequently achieves top performance in structured data competitions. When processed data assumes tabular format with rows representing field-date combinations and columns containing features such as temperature and soil moisture, XGBoost typically delivers optimal performance through its efficiency, accuracy, and robustness.

PyTorch: This deep learning framework offers flexibility through dynamic computation graphs, making it suitable for research and complex architectures. Implementation would be considered under specific circumstances:

Direct processing of spatial data through Convolutional Neural Networks for automated feature extraction from raw satellite imagery

Modeling strong temporal dependencies using Recurrent Neural Networks or Transformers for extended field time series
For initial project phases, Scikit-learn and XGBoost provide sufficient capability, while PyTorch represents an advanced option for capturing complex spatio-temporal patterns beyond simpler models' capacities. **(Optional Component)**

III. Data Handling and Geospatial Processing

Pandas: This library forms the foundation for data manipulation through DataFrames, enabling structured tabular data operations. Applications include loading CSV files, data cleaning, merging heterogeneous datasets, and feature engineering.

GeoPandas: Extending Pandas with geospatial capabilities, this library incorporates geometry columns containing shapes such as points, lines, and polygons. Essential functions include:

Reading agricultural field boundaries within the Ebro Basin
Performing spatial operations including proximity analysis between fields and weather stations
Conducting zonal statistics through integration with rasterio to calculate average vegetation indices or soil moisture within field polygons

NetCDF4: This library facilitates interaction with NetCDF files, the standard format for multidimensional climate and meteorological data. Copernicus program datasets, particularly ERA5 reanalysis weather data, are distributed in this format. The library enables extraction of time series for variables including temperature, precipitation, and wind speed across specified geographic regions.

Rasterio: Specialized for geospatial raster data processing, this library supports:
Accessing satellite imagery from Sentinel-2 (vegetation indices) and Sentinel-1 (soil moisture)
Reading pixel values with associated geographic coordinates
Clipping raster data to specific geographic boundaries
Integrating with GeoPandas for zonal statistics computation across field polygons

IV. Visualization

Matplotlib: As the foundational plotting library for Python, this tool enables creation of static, publication-quality visualizations including soil moisture time series, precipitation histograms, and regional maps.

Plotly: This modern graphing library provides interactive capabilities including zoom, pan, and hover data display. It is particularly suited for developing final dashboard interfaces, enabling creation of field maps color-coded by irrigation recommendations with detailed time-series visualizations accessible through user interaction.