demeter

TITLE: Artificial Intelligence-based Decision Support Components

DATE: December 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857202.

AI-based Decision Support Tools

1 Summary

DEMETER aims to lead the Digital Transformation of the European Agrifood sector based on the rapid adoption of advanced technologies, such as Internet of Things, Artificial Intelligence, Big Data, Decision Support (DSS), Benchmarking, Earth Observation, etc., to increase performance in multiple aspects of farming operations, as well as to ensure the long-term viability and sustainability of the sector. It aims to put these digital technologies at the service of farmers using a human-in-the-loop approach that constantly focuses on mixing human knowledge and expertise with digital information.

To enable the achievement of those objectives, and to promote the targeted technological, business, adoption and socio-economic impacts, DEMETER is designing and developing targeted decision support systems to enable the delivery of tailored advisory services to the agricultural sector. These DSS services combine data analytics, AI-based expert systems, and machine learning techniques to provide precision decision support to the users and provide the intelligence to several DSS components which cover different situations (e.g., crop identification, irrigation needs or animal care). These should help the farmer with understanding the current state and, eventually, predicting future states.

The AI modules developed, adapted and/or deployed will help in solving the needs from the different pilots at the same time that they are integrated with the DSS components. These AI modules have been developed as generic as possible to allow their application in other pilots' sites whenever possible.

2 Artificial Intelligence Based Components

2.1 AI Model for Component 4.A.1: Plant Yield Estimation

The yield prediction model uses a smoothed Sentinel-2 timeseries (daily Normalised Difference Vegetation Index, NDVI, values) to predict potato yields on field level using data from harvesting machines, produced by DEMETER partner AVR, as ground truth data to train the model. The same process could be used to train a model for other crops.

The training process starts with the generation of the crop growth curve. From an AIM description of the field (crop type, polygon, planting date), a curve can be

retrieved from a Sentinel-2 timeseries service, with data for cloudy days interpolated using a smoothing algorithm. Data is generated from starting date to the date on which the prediction is done (2-3 weeks before harvest).

To predict the yield, the smoothed timeseries is fed into the regression model, an ensemble of 10 Multi-Layer Perceptron (MLP) with each MLP being a 3-layer neural network with 1 output neuron representing a scaled-down value of the predicted yield. The output of the regression model is then unscaled using a fixed scaler valid only for storage potatoes.

2.2 AI Model for Component 4.A.2: Plant Phenology Estimation

This model uses a BBCH scale, which represents the phenology state (winter buds, flowering etc.) of plants numerically. The service has been pre-trained for olive plants using Random Forest model and the connection to the Copernicus API.

The Copernicus service is used to extract the temperature measurements from Meteostat weather API for the user-supplied geographical coordinates (latitude and longitude) and date.

The simplest prediction model uses the Allen formula to compute the Growing Degree Day (GDD) from maximum and minimum temperature measurements registered since January 1st. The pre-trained random forest (RF) model is applied to the GDD and the user-selected Day Of Year (DOY) resulting in a rescaled BBCH representing a specific olive phenology state.

A more complex model using an extra-tree regressor model with input from Copernicus ERA5 climate data is also being produced.

2.3 AI Model for Component 4.A.4: Crop Type Detection

The purpose of this AI model is to detect the crop type for a given polygon and a given timeframe (growing season of the crop), using satellite data as input. A Recurrent Neural Network using the TensorFlow5 deep learning framework has been used to work on a timeseries of images as the differences in the timeseries is key to good identification.

The input to the recurrent neural network is a combination of Sentinel-1 and Sentinel-2 timeseries at the field-polygon level, with fixed start and end of the dates (growing season). The timeseries contain information from all Sentinel-2 bands, augmented with the NDVI and Sentinel-1 Vertical (VV) and Horizontal (VH) polarisation signals, and both for Descending and Ascending orbits.

The 2 timeseries are resampled to 5-day intervals and fed into separate stacked LSTM networks (Long-Short Term Memory, a specific type of recurrent neural network). The outputs of both LSTM stacks are concatenated in the second layer of the network. The final layer (output layer) is a classifier implemented as a dense neural network, with the number of output neurons equal to the number of crop types we wish to detect.

2.4 AI Model for Component 4.B.2: Reference Evapotranspiration Prediction

This algorithm combines timeseries model predictions using ML with the Reference Evapotranspiration (ET_0) calculated using the Penman-Monteith method with weather predictions from several meteorological services to obtain the final forecast value that can be used to schedule a dynamic irrigation plan.

The algorithm is run daily with each implemented timeseries prediction model (TSM) storing its ET_0 predictions for the next days in a database. The most suitable TSM is decided by comparing most recent day's stored data (ET_0 predictions in previous execution) with the most recent day's real ET_0 calculation (Penman-Monteith method) selecting the one that least underestimates it.

2.5 AI Model for Component 4.B.3: Soil Moisture Estimation

The data provided by in-field soil moisture probes (one-point data) is used to generate a model capable of quantifying the amount of water on the plot surface (2D data), complementing the soil probes information. An optical trapezoidal model (OPTRAM) ML algorithm has been implemented that is fed by both soil moisture probe data and Sentinel-2 multispectral images. The output of this model can then be integrated into the irrigation model to estimate the amount of irrigation water that should be supplied to the crop to keep it at field capacity.

This model is built in a plot by the following steps:

- Collect Sentinel-2 multispectral images and the corresponding data from the ground soil moisture probes for as long a period as possible.
- Build the Shortwave-infrared Transformed Reflectance (STR) and NDVI space from the Sentinel-2 images.
- Fit the wet (STRw) and dry edges (STRd) of the previous space.
- Calculate the normalised soil moisture content $W = f(STR, STRd, STRw)$ for every available pixel in the plot, including those for the ground probes.
- Perform linear regression analysis to obtain the soil constants, minimum dry (ϑd) and maximum wet soil moisture (ϑw).

- Compute surface soil moisture $\vartheta = f(W, \vartheta d, \vartheta w)$ for the ground truth datapoints.

2.6 AI Model for Component 4.B.4: Crop Water Status Anomalies Detection

This model identifies possible anomalies in the crop extension (2D) related to the plants' water status using multispectral analysis of images provided by Sentinel-2 satellite. Images of the crop over several seasons are compared with the latest image obtained to classify the pixels according to the expected behaviour extracted from the historic data of the same crop in the same or adjacent plots.

A Z-score map is used to determine the anomalies by measuring the number of standard deviations a point is away from the mean. The Z-score for each of the pixels of the most recent image can be calculated by comparing it to the corresponding pixels of previous seasons, using the formula:

$$NDVI_{Z-score} = \frac{NDVI_{Z-score} - \mu_{NDVI_{1\ldots n-1}}}{\sigma_{NDVI_{1\ldots n-1}}}$$

2.7 AI Model for Component 4.E.1: Pest Estimation with Sterile Fruit Flies

This model makes use of a generic, Neural Networks-based, counting component which is being trained to count the numbers of sterile and non-sterile fruit flies from images. The images are taken under ultraviolet light as a fluorescent dye has been applied to the sterile fruit flies.

Training was performed using labelled images, some of which were produced under lab conditions. The labelling being performed manually by experts.

2.8 AI Model for Component 4.E.2: Estimate Temperature Related Events

This AI model is based on the same technology as 4.A.2 but rather than phenology events, it predicts the development stages of pests based on weather conditions. It has been pre-trained using data collected by monitoring olive fruit fly activity. The weather data is collected from the Copernicus ERA5 data service.

2.9 AI Model for Component 4.F.1: Estimate Milk Production

This model develops individual, cow-specific lactation curves as a basis for predicting future milk yield. The milk yields are predicted on a 15-day interval over a total of 345 days. The input data is a combination of previous milk yields, over the same periods and intervals as the output, and a variety of variables including the breed, calving number, time of year, feed consumption and days in milk.

The algorithm used is named CatBoost; a well-established gradient boosting regressor with special support for categorical features which combines lots of small decision trees into one big model. The model is trained tree-by-tree and therefore allows the selecting of new trees and the values in each tree to be dependent on previous trees. The selecting of weights on the base trees is done by gradient descent on the loss function. The loss function is a MultiRMSE, the mean RMSE for each prediction (23 steps in this model).

The hyperparameter tuning was done using grid-search on the parameters: 'depth', 'bagging_temperature', 'l2_leaf_reg', 'learning_rate', and 'grow_policy'.

The best score is chosen by using the Mean Absolute Cumulative Error:

$$score = \frac{\sum_{i=1}^{N} \left| \sum_{d=1}^{dim} \left(a_{i,d} - t_{i,d} \right) \right|}{N}$$

where:

- N is the number to predict (rows in dataset),
- dim is the number of timesteps ahead to predicted (columns in dataset),
- a is the true value and t is the predicted.

2.10 AI Model for Component 4.G.1: Estimate Animal Welfare Condition

The ML algorithm on Animal Welfare evaluates of the health status of the cows analysed to determine the degree of well-being in terms of nutrition, hygiene, rest, and movement and, consequently, to evaluate their productivity. A Random Forest algorithm is being used.

The training data contains fat/protein ratio, the electrical conductivity, the total days at rest and the total daily rest, accompanied by the health status by pathology (ketosis, mastitis, and lameness) assigned "manually" by the farmer. By processing this flow, the Random Forest algorithm learns in what circumstances a cow is healthy and in which circumstances it is sick:

- If the cow rests too little or too much, it may get too tired and limp or may already be lame and have difficulty getting up.
- A diet which is too unbalanced on fats, can favour the onset of ketosis.
- If the electrical conductivity of the milk taken exceeds a certain threshold, it is very likely that the cow is suffering from mastitis.

The values calculated by the algorithm can be compared to the actual health status values assigned manually. When there is a high degree of match between the two

values, the algorithm can be used to make predictions about the animal health and hence help to identify ways to improve the quality of life of the cows.

2.11 AI Model for Component 4.G.2: Poultry Well-Being

The algorithm for poultry well-being aims to provide additional support to an existing platform for poultry monitoring, which is based on micro-electronics devices and to integrate the algorithm into a smart cloud-based system.

The extension is video-based detection of chicken well-being - the main parameters are classified chicken size and the chicken movements on farms to facilitate improvement of the breeding process. The AI algorithm is much more complex and requires careful deployment of edge devices to capture representative datasets, existing devices with video capturing capability, edge data processing and machine learning. This algorithm runs directly on the edge device (i.e., the camera) and estimates the weight and activity from the images. The annotated dataset is compiled from 3500 images captured over 35 days. The algorithm is based on round estimation of the area that represents the shape of each chicken. The algorithm was trained by using sets of less than 1000 images for chicken recognition and 150 images for weight estimation.

2.12 AI Model for Component 4.H.1: Milk Quality Prediction

This ML algorithm focuses on the ability to predict milk quality by analysing data collected during the lactations of dairy cows. To allow the algorithm to “understand with some certainty and reliability” whether the analysed sample refers to high- or low-quality milk, the training data must be qualitatively representative.

Fourier-transform infrared spectroscopy (FTIR) analysis was used for the extraction of milk quality data, to obtain reliable data for both the training and prediction levels. The training data stream contains the data for the FTIR analyses with the quality grade assigned "manually". The Random Forest algorithm then learns in what circumstances the milk is of high, medium, or low quality. The parameters analysed are caseins, density, fats, proteins, cryoscopic point, lactose, and urea. If the milk is not very dense and the cryoscopic point is too high, the percentage of water present in the milk is probably excessive indicating a lower quality. Similarly, fats and proteins indicate the genuineness and nutritional value of milk. Too much uric acid indicates that the cow is on the wrong diet and is probably consuming too much protein.

3 Conclusions

The AI modules developed, adapted and/or deployed will help in solving the needs from the different pilots at the same time that they are integrated with the DSS components. These AI modules have been developed as generic as possible to allow their application in other pilots' sites and in farms not directly involved in DEMETER whenever possible.

4 Annex 1 – Terminology

AIM: The DEMETER Agriculture Information Model is a data model and ontology that describes all data needed by DEMETER applications and the usage of which ensures semantic interoperability between data and various components.

Allen method is a modified sine wave method for calculating degree days.

BBCH: The abbreviation BBCH derives from the names of the originally participating stakeholders: "**B**iologische **B**undesanstalt, Bundessortenamt und **CH**emische Industrie". The **BBCH-scale** is used to identify the phenological development stages of plants using a decimal code system, which is divided into principal and secondary growth stages.

CatBoost is a high-performance open-source library for gradient boosting on decision trees.

The **Copernicus API** provides access to the Copernicus Climate Data Store.

Deep learning is a type of machine learning and artificial intelligence (AI) that imitates the way humans gain certain types of knowledge. While traditional machine learning algorithms are linear, deep learning algorithms are stacked in a hierarchy of increasing complexity and abstraction.

Descending and Ascending Orbits: All SAR satellites travel from the north pole towards the south pole for half of their trajectory. This direction is referred to as their descending orbit. Conversely, when satellites travel from the south towards the north pole, it is said to be in an ascending orbit.

Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. An extra-trees regressor - this class implements a meta estimator that fits a number of randomized decision trees (a.k.a.

extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

ERA5 is the fifth generation ECMWF atmospheric reanalysis of the global climate covering the period from January 1950 to present. ERA5 is produced by the Copernicus Climate Change Service (C3S) at ECMWF. ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables.

Evapotranspiration can be defined as the sum of all forms of evaporation plus transpiration, but here, it is the sum of evaporation from the land surface plus transpiration from plants.

Fourier-transform infrared spectroscopy (FTIR) is a technique used to obtain an infrared spectrum of absorption or emission of a solid, liquid or gas. An FTIR spectrometer simultaneously collects high-resolution spectral data over a wide spectral range. The term Fourier-transform infrared spectroscopy originates from the fact that a Fourier transform (a mathematical process) is required to convert the raw data into the actual spectrum.

Gradient boosting is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function.

A **loss function** is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event.

Meteostat is one of the largest vendors of open weather and climate data.

A **Multi-Layer Perceptron** is a class of neural network involving at least 3 layers of perceptron which can decide whether or not an input belongs to a specific class.

The **OPtical TRApezoid Model (OPTRAM)** relies on a physical linear relationship between the soil moisture content and shortwave infrared transformed reflectance (STR).

Polarisation Signals:

VV - for vertical transmit and vertical receive (VV)
HV - for horizontal transmit and vertical receive (HV)
VH - for vertical transmit and horizontal receive (VH).

Random forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

Recurrent neural networks (RNN) are a class of neural networks that are helpful in modeling sequence data. Derived from feedforward networks, RNNs exhibit similar behavior to how human brains function. Simply put recurrent neural networks produce predictive results in sequential data.

Regression model is an algorithm for estimating the relationship between a dependent outcome variable and one or more independent variables. The simplest algorithms is linear regression which finds the line which most closely fits the data according to a specific mathematical criterion.

Root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model, or an estimator and the values observed

Sentinel 1&2: The Sentinel missions from the European Space Agency's Copernicus Programme comprise several satellite constellations for Earth observation. Sentinel -1 has two polar-orbiting satellites providing radar imagery. Sentinel-2 has currently two polar-orbiting satellites providing multispectral imagery in 13 bands covering the Visible, Near InfraRed, and Short-Wave InfraRed parts of the electromagnetic spectrum.

TensorFlow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks

Z-scores are multiples of standard deviations from the mean of a normally distributed population.

demeter

CONTACT US

Kevin Doolin, Project Coordinator

Kevin.Doolin@WaltonInstitute.ie

or email INFO@H2020-DEMETER.EU

For more information visit **WWW.H2020-DEMETER.EU** and follow us on social media:

H2020DEMETER **H2020DEMETER** **h2020-demeter** **h2020-demeter**