Machine Learning techniques for Hydroponic Cultures

Leonardo Plini*[1], Davide Mascolo[1]

[1] Sapienza University of Rome
plini.2000543@studenti.uniroma1.it; mascolo.2001991@studenti.uniroma1.it

Abstract: Hydroponics is an innovative agricultural technique that enables the cultivation of plants without the use of soil, providing controlled conditions for optimal plant growth. In recent years, machine learning (ML) techniques have gained prominence in various domains, including agriculture, due to their ability to analyze large datasets and derive valuable insights: the combination of ML and the opportunity to control all the inputs in an hydroponic cultivation represents an invaluable chance to reduce resource requirements and increase the production in line with the constraints imposed by climate change. In this work, we tested four different machine learning models, namely, random forest (RF), support vector machine (SVM), extreme gradient boosting (XGB) and a neural network. These models were tested in two different scenarios considering two sets of variables. The first scenario is done considering all the features of the dataset while the second scenario is characterized only by the features that can be measured during the cultivation. The best result is obtained in the second scenario with extreme gradient boosting (XGB) that achieved a value of 8.37 for mean absolute error (MAE), 8.20 for mean bias error (MBE) and 13.16 for root mean square error (RMSE).

Keywords: Hydroponic Cultures, Machine Learning, innovative agricultural techniques.

[*] Corresponding Author: plini.2000543@studenti.uniroma1.it

1. Introduction

The climate change represents a urgent and complex problem that can impact human life in different ways, but the main challenge is related to food supply due to land and water scarcity. According to the Organisation for Economic Cooperation and Development (OECD), farming accounts for around 70% of water used in the world and contributes to water pollution from excess nutrients, pesticides and other pollutants.

For this reason, sustainable management of water in agriculture becomes a critical challenge to address water scarcity and a global focus for every country. In this context, technology has made important steps proposing new frameworks that improve the current usage of resources such as hydroponic, aeroponic and aquaponic systems. All these apparatuses require careful implementations, and several modelling decisions must be taken into consideration to reach the best performance in terms of production.

In this landscape, Machine Learning techniques are acquiring a central role. In fact, we can leverage these methodologies to clarify the effect of each component in a controlled environment.

The purpose of this paper is to analyse the data of a tomato crop in an hydroponic setup to highlight the possible benefits of this procedure and the plausible extension to other cultivations.

2. Literature Review and Related works

Technology has always played a central role when it comes to agricultural applications and it has driven huge changes during the history. However, only in the recent years has been possible to leverage internet applications and huge compute resources to gain relevant insights. Impedovo e at. [Impedovo et al., 2018] showed how to manage heterogeneous information and data coming from real dataset that collect physical, bi-
ological and sensory values. Signore et al [Signore et al., 2018] designed and implemented an experiment in ”la Noria” Farm of the Institute of Science of Food Production of the National Research Council using the Nutrient Film Technique (NFT) for an hybrid variety of cherry tomato. The data were public and accessible through the portal Mendeley Data and they were used after a careful manipulation for this study.

Please refer to the previous paper for a detailed description of the setup.

Signore et al in [Signore et al] leveraged the same data to describe a precise management of the nutrient solution that allowed discarding a lesser amount of water and nutrients into the environment, improving the sustainability of the crop in a Mediterranean environment. Meshram et al. in [Meshram et al., 2021] presented an extended survey on the latest machine learning application in agriculture to alleviate the problems in the tree areas of pre-harvesting, harvesting and post-harvesting.

El-Ssawy , Al-Anasari e al. in [Mokhtar et al., 2021] applied machine learning models to an hydroponically grown lettuce yield and serves as a point of reference for our study.

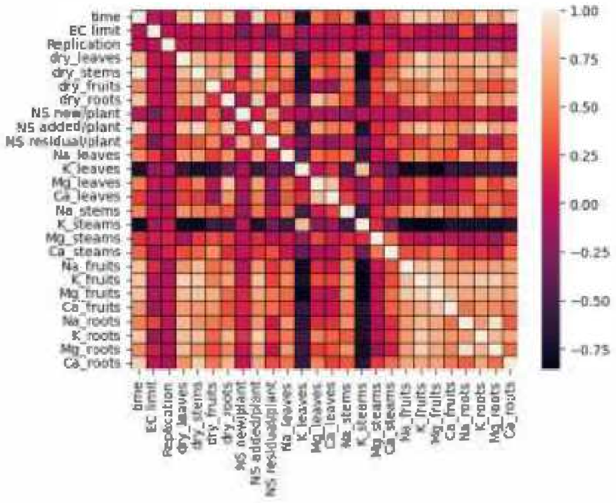


Figure 1 – Correlation Matrix

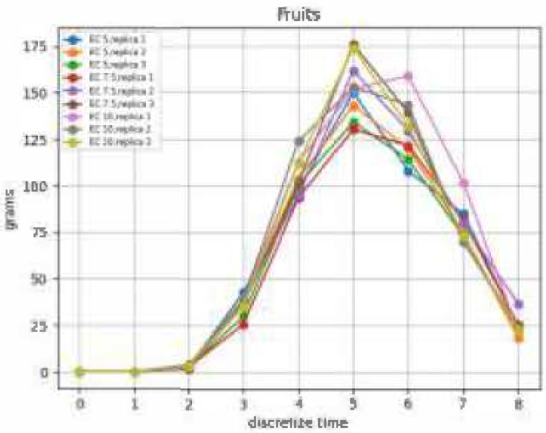


Figure 2 – Dry Fruit curves

3. Proposed Methodology

The data went through a pre-processing phase to aggregate all the relevant information considering a discrete time period representation and summing all the quantities within the relative time window. A brief initial data analysis allowed to gain relevant insight on the data.

Considering the correlation matrix in Figure 1, we can notice how the weight of the fruit is related to the period and to the Potassium, in line with the scientific information on this crop that suggests that this nutrient has a positive impact on both the size and the color of the fruit [Weinert et al., 2021]. Two scenarios were taken into account to compare the performances of the methods where a different number of independent variables is considered. In particular, the second scenario we omitted all the quantities related to the fruits in order to consider a more realistic scenario.

We compare the performances using three distinct metrics:

- Mean Absolute Error (MAE)= $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- Mean Bias Error (MBE) = $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$
- Root Mean Square Error (RMSE) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

The data were splitted into train,validation and test with the proportion 80%, 10% and 10% considering a random time instant for each sample in both test and validation.

The analysis considered Random Forest, Support Vecotr Machine, Extreme Gradient Boosting and a Deep neural Netowork in both scenarios. Each framework was optimized by means of a grid search to ensure an optimal choice of the parameters and hyperparameters.

3.1 Machine Learning Models descriptions

3.1.1 Support Vector Machine

The Support Vector Machine (SVM) algorithm offers significant results with minimal computational requirements.

In SVM, every data record represents a point in a feature space with p dimensions, where p is the number of features.

When p is 2 like in Figure 3, SVM identifies a line that maximizes the distance from the nearest point in each category to this line. However, when p is greater than 2, SVM finds a hyperplane that optimally separates the data points belonging to each category, known as the Maximum Margin Hyperplane (MMH) in the sense that it minimizes the error.

The data points closest to the MMH, called Support Vectors (SV), define the MMH and serve as a compact representation of the model. Each category must have at least one SV, but it can have multiple SVs. By mapping a new data record onto the corresponding region, SVM predicts its category, effectively combining elements of nearest neighbour and regression methods. The amalgamation of these techniques in SVM enables the modelling of intricate data relationships.

3.1.2 Support Vector Machine

Random Forest (RF) is an enhanced version of a decision tree algorithm that combines the fundamental principles of bagging with random feature selection to introduce additional diversity to the decision tree models. Decision tree learners are robust predictive models that employ a tree- like structure to establish relationships between features and outcomes. This structure resembles a tree starting with a broad trunk and branching into narrower branches as it extends upwards.

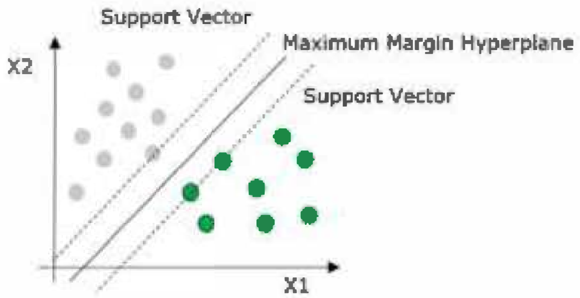


Figure 3 - Support Vector Machine (from UNECE, ”Learning for Official Statistics”, 2021

Similarly, a decision tree learner utilizes branching decisions to guide examples towards a final predicted class value. While a decision tree is built on the entire dataset, incorporating all relevant features, RF randomly selects observations and specific features to construct multiple decision trees, subsequently averaging the results for making predictions. In the RF model, the Gini Coefficient is employed. The Gini coefficient determines how nodes branch in a decision tree and is calculated as follows:

$$Gini = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}} \tag{1}$$

where ”n” represents the number of observations. Similarly, entropy is another indicator that governs node branching in a decision tree and is calculated as:

$$E_{Split} = \frac{N_1}{N}E_1 + \frac{N_2}{N}E_2 \tag{2}$$

where N_1 and N_2 are the numbers of items in each set after the split, and E_1 and E_2 are their corresponding entropies.

Random Forest offers several advantages over other machine learning algorithms. It selectively chooses essential features and can be effectively applied to datasets with an exceptionally large number of features. Figure 4 illustrates a schematic diagram of the RF model. The final predicted value in the RF model is obtained by averaging the predictions from all individual trees.

3.1.3. Extreme Gradient Boosting

Introduced in 2016, eXtreme Gradient Boosting (XGB) has gained immense popularity as a machine learning technique. Boosting, the core concept behind XGB, involves adding new models to the ensemble in a sequential manner.

This technique enhances the bias-variance tradeoff by initially employing a weak model and progressively improving its performance by constructing new trees. Each subsequent tree aims to address the most significant errors made by the previous one. XGB represents a noteworthy advancement in Gradient Boosting. Figure 5 illustrates the schematic diagram of the XGB model.

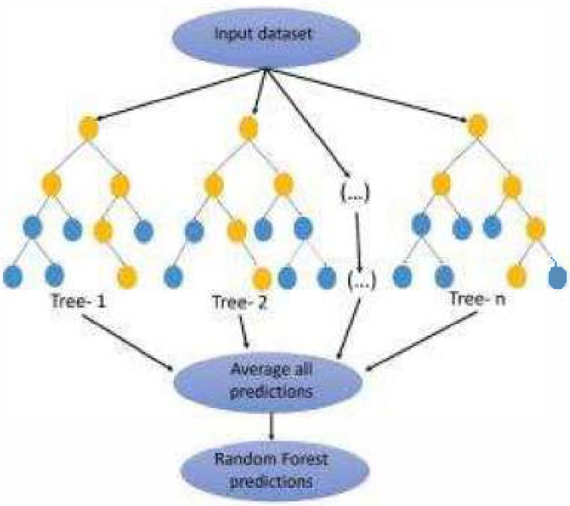


Figure 4: Random Forest diagram (from Sahour et a., ”Random forest and extreme gradient boosting algorithms for streamfow modeling using vessel features and tree-rings”, 2021.

The process of gradient boosting begins with a set of predictors (X_1,...,X_n) used to predict the corresponding target values (Y_1,...,Y_n). We train a model F(X) → Y and minimize the sum of the loss function $J = \sum_{i=1}^{n} L(Y_i, F(X_i))$, to enhance the model F(X). The loss function L measures the difference between the prediction F(X) and the target Y and is a differentiable convex function. The following iterations are performed: firstly, we calculate the negative gradients of J with respect to F(X_i), denoted as $-\frac{\partial J}{\partial F(X_i)}$. Next, a regression tree h is fitted to the negative gradients $-\frac{\partial J}{\partial F(X_i)}$. Finally, F(Xi) is updated by adding γh, where γ is the step size used to approach the estimated minimum of J. This iterative process continues until the desired accuracy is achieved. In XGB, the loss function is defined as follows:

$$J = \sum_{i=1}^{n} L(Y_i, F(X_i)) + \Omega(h) \tag{3}$$

where $\Omega(h) = \gamma T + \frac{1}{2}\lambda||\omega||^2$. Here, T represents the number of leaves in the tree, and ω corresponds to the leaf weights.

Variable Importance (VI) quantifies the statistical significance of each variable in the data concerning its impact on the generated model. It represents the ranking of each predictor based on its contribution to the model. The calculation of variable importance involves measuring the decrease in error when the data is split by a particular variable. Subsequently, the relative importance is determined by dividing the variable importance by the highest value of variable importance, ensuring that the values are confined within the range of 0 and 1. In tree-based regression models such as Random Forest (RF) and XGB, the measure of VI is determined by the frequency of a variable being selected for splitting and the extent to which the model improves as a result of the split.

3.1.3. Extreme Gradient Boosting

An Artificial Neural Network (ANN) or Deep Neural Networks (DNN) comprises interconnected nodes known as ”artificial neurons” that aim to mimic the neurons found in the human brain. These neurons are usually organized into layers, with each layer potentially performing distinct transformations on its inputs. The flow of signals starts from the initial layer, known as the input layer, and propagates through the subsequent layers, eventually reaching the final layer, referred to as the output layer. In some cases, the signals may traverse through the layers multiple times during the network’s processing.

Each node within the network receives a set of m inputs, denoted as xi, which undergoes processing through the application of weights wi and the addition of a bias term. This computation is performed using the equation $s = (\sum_{i=1} mw_ix_i) + bias$ =. Subsequently, an activation function φ is applied to the resulting value s, producing the output of the neuron, denoted as $y = \varphi(s)$. This output is then transmitted to other nodes within the network.

Depending on the specific application, various types of activation functions can be utilized, resulting in discrete or continuous outputs that may be bounded or unbounded. In our experiment we used the ReLu which is definited as:

$$ReLu(y) = \begin{cases} y, & \text{if } y \geq \bullet \\ 0, & \text{if } y < \bullet \end{cases} \tag{4}$$

Several advanced machine learning algorithms are founded on the principles of DNNs, showcasing their effectiveness in learning complex patterns within datasets. However, it is important to note that larger DNNs often require significant computational power, necessitating investments in hardware such as CPUs or GPUs.

A Multi-Layer Perceptron (MLP) is a type of feedforward Artificial Neural Network (ANN) that includes one or more hidden layers. In an MLP, signals flow from the input layer through the hidden layers and ultimately reach the output layer, hence the term ”feedforward”. Each node, except for the input nodes, functions as a neuron, while each input node represents a distinct feature of the dataset. The input value of a feature is propagated to the first node, which combines and processes the received inputs and applies an activation function to transform the result. This transformed output is then passed to the subsequent layer.

To ensure effective performance, MLPs need to be trained using a training set where the desired outputs are known. The training process involves iteratively adjusting the weights used by the artificial neurons, typically through a technique called ”Backpropagation”. Backpropagation aims to minimize a loss function, such as the average squared difference between the predicted and actual outputs, by updating the weights. This iterative adjustment of weights is often carried out using a gradient descent method, which makes incremental changes based on the results obtained from each successive data point in the training set. The learning rate, a hyperparameter of MLP, controls the speed of the weight updates.

The complexity of an MLP determines its ability to recognize complex relationships within the data. However, it’s important to note that as the MLP becomes more intricate, the computational requirements increase significantly due to the number of layers and neurons in each layer.

4. Results and discussion

Given the numerosity of the sample, the results come with no surprise: a Neural Network is not suited for this dataset due to the large number of parameters that typically require more examples to reach competitive performances. On the other hand, Random Forest and Extreme Gradient Boosting have shown better performances with respect to Support Vector Machine. In this case, Extreme Gradient Boosting is the best method on average, reaching very low errors in scenario 2 as the tables shows.

It worth noticing that the importance assigned to the variable" time" (Figure 7) was considered the most influential one in both scenarios the XGB and the variable "EC limit", which gives an indication of the salinity level of the water used, was not considered a central feature. This last result confirms the resistance of certain hydroponic plants to different levels of salinity in accordance with the literature.

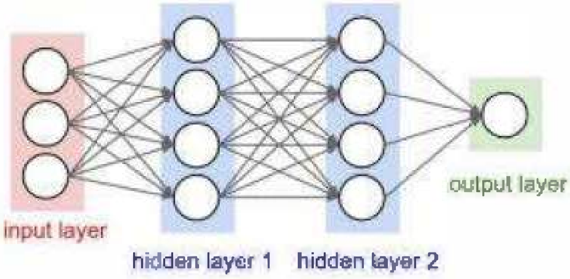
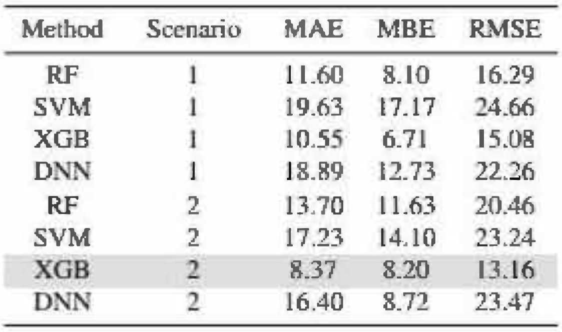


Figure 6: Deep Neural Network, image from Dashanka Nadeeshan De Silva , Multilayer Perceptron with Pytorch Github Repository



Method	Scenario	MAE	MBE	RMSE
RF	1	11.60	8.10	16.29
SVM	1	19.63	17.17	24.66
XGB	1	10.55	6.71	15.08
DNN	1	18.89	12.73	22.26
RF	2	13.70	11.63	20.46
SVM	2	17.23	14.10	23.24
XGB	2	8.37	8.20	13.16
DNN	2	16.40	8.72	23.47

Table 1: Results of the Models

It is also important to state that the reduced numerosity of the dataset does not allow to draw clear conclusions on both the analysis and both the analysis and the quantification of the errors might be taken into account using a larger dataset to provide some confidence levels on the results.

5. Conclusion and Future Works

The study shows how different machine learning techniques can be employed to exploit the information stored in agricultural data and in particular for hydroponic setups. The usage of advanced machine learning techniques rely on the availability of large amount of data that are very expensive for this kind of experiments. For this reason, the research require the active participation of both state institutions and private firms to maximize the diffusion of the data and ensure a concrete pervasiveness, considering that food supply represents a central asset for each nation. DNN could be used for each plant variety to address specific problems related to the management of the Nutrient solution, the regulation of the temperature and all the other parameters the directly affects the growth. Future works might consider the introduction of robotic systems based on Artificial Intelligence algorithm integrating them into the process of production.

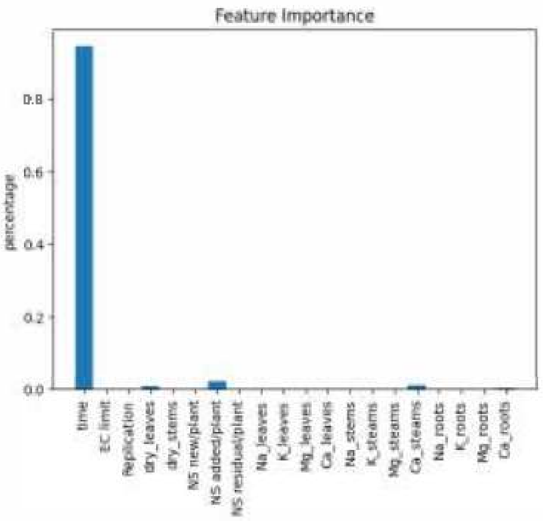


Figure 7: Variables Importance

References

Balducci F., Impedovo D., Pirlo G., ”Machine Learning Applications on Agricultural Dataset for Smart Farm Enhancement”, 2018.

Signore A., Serio F., Santamaria P., ” A Targeted Management of Nutrient Solution in a Soilless Tomato Crop According to Plant Needs”, 2016.

Signore A., Serio F., Santamaria P., ”Growth Analysis and Nutrient Solution Management of Soil-less Tomato Crop in a Mediterranean Environment.

Meshram V., Patil K., Meshram V., Hanchate D., Ramkteke S.D.,” Machine learing in agriculture domain: A state-of-art survey”,2021.

Mokhtar A., El-Ssawy W., He H., Al-Anasari N., Sammen S. S., Gyasi-

Agyei Y., Abuarab M.,” Using Machine Learning Models to Predict Hydroponically Grown Lettuce Yield”, in 2022.

Weinert C.H., Sonntag F., Egert B., Pawelzik E., Kulling S. E. , Smit I., ”The effect of potassium fertilization on the metabolite profile of tomato fruit (Solanum lycopersicum L.)” ,2021.