Improve Irrigation Timing Decision for Agriculture using Real Time Data and Machine Learning

João Cardoso
Department of Science, Technology and Information
ISCTE - Instituto Universitário de Lisboa
Lisbon, Portugal
Email: jmbco1@iscte-iul.pt

André Glória
Department of Science, Technology and Information
ISCTE - Instituto Universitário de Lisboa
Lisbon, Portugal
Email: afxga@iscte-iul.pt

Pedro Sebastião
Department of Science, Technology and Information
ISCTE - Instituto Universitário de Lisboa
Lisbon, Portugal
Email: pedro.sebastiao@iscte-iul.pt

***Abstract*—With the constant evolution of technology and the constant appearance of new solutions that, when combined, manage to achieve sustainability, the exploration of these systems is increasingly a path to take. This paper presents a study of machine learning algorithms with the objective of predicting the most suitable time of day for water administration to an agricultural field. With the use of a high amount of data previously collected through a Wireless Sensors Network (WSN) spread in an agricultural field it becomes possible to explore technologies that allow to predict the best time for water management in order to eliminate the scheduled irrigation that often leads to the waste of water being the main objective of the system to save this same natural resource.**

***Index Terms*—Machine Learning, Neural Network, Decision Tree, Support Vector Machine, XGBoost, Random Forest, Sustainability, Smart Irrigation**

I. Introduction

Agriculture has always been the main supplier of food for society, being responsible for more than 74% [1] of the population daily consumption. In order to keep these production values and evolving them to meet the needs of the increasing population, water consumption has been the main concern. In this topic, resources to technological solutions have demonstrate to be the best bet to meet the needs.

Regarding the issue of water consumption, its application continues to be mostly controlled by human means, leading many times to mismanagement, higher than necessary consumption and consequently to higher financial expenses and the waste of this natural resource.

Technological solutions have been increasingly concerned with monitoring the fields in order to understand the best times for their harvest and controlling their values throughout the season. This fact makes possible a more in-depth research in order to predict the best water management heights, leading to less waste and an improvement in harvests.

Thus, the main objective of this paper is to understand the possibility to predict, automatically, using machine learning solutions, the best time of day for irrigation, based on local sensor and weather data.

This work was supported in part by ISCTE - Instituto Universitário de Lisboa from Portugal under the project ISCTE-IUL-ISTA-BM-2018

So, and based on related work in the area of agricultural land monitoring, it is possible to make the correct analysis of which data are the most important to take into account when developing the algorithms [2].

The system already implemented by the authors in [2] will be the basis of the research for this paper. This system uses a wide range of sensors that are strategically spread over the agricultural fields in order to collect the data needed for correct monitoring. Then this data is sent, using Wireless Sensor Network (WSN), to a central server, where a database is hosted, in order to be able to store the data collected over long periods of time. This high number of data will allow the correct development and training of the algorithms to accomplish the proposed goal of this paper, understating which is the best machine learning algorithm to predict the ideal irrigation hours.

According to the results obtained by the authors in [3], it is possible to reach savings up to 40% in water. These water saving results are obtained only through the study of formulas that calculate the amount of water that needs to be administered to the fields. So, the implementation of machine learning algorithms, as this paper will study, is in a good position to achieve even better savings results of this natural resource.

This article presents a study of Machine Learning Classification algorithms where Random Forest (RF), Neural Networks (NN), XGBoost, Decision Trees (DT) and Support Vector Machine (SVM) will be studied.

In addition, and in order to be able to study correctly the algorithms already mentioned, a dataset and a methodology were created in order to enable the correct study, leading to the best possible optimization. The whole process of creating the spoken dataset and the methodology followed will be explained later. Later will be presented the conclusions drawn from the study present in this paper.

II. Related Work

With the evolution of technology and the constant development of solutions in the area of the Internet of Things (IoT) in parallel with intelligent solutions produced in the area of

artificial intelligence and machine learning, multiple solutions have already been developed with the aim of combining the two in order to obtain cheaper solutions and with the purpose of saving natural resources.

In the study made in [4], the author developed a system that applies artificial neural network techniques "for water level prediction, a fuzzy logic control algorithm for sluice gate setting period estimation, and hydraulics equations for sluice gate level adjusting". As input for the pretended prediction, the author only gives to the algorithm a dataset with the last three days of water level, being this a small amount of data since the main goal is to reach the most exact prediction possible being the amount of data given as input a concern.

Regarding to the study made by the author of [5], this article presents a study of machine learning applications in agricultural supply chains. Although the author have not developed any system or script, had conclude, through a heavy research that the most explored algorithms are neural Networks, being these the most used for agricultural solutions.

In [6], the authors present, through the collection of land data using a WSN, a study where machine learning algorithms (SVM and RF) are applied in order to understand the irrigation needs of the land under study, with an accuracy in the order of 80%. Although the study developed in [6] has an intensive research on the level of collected data, and there has been an investment in the analysis of the values through formulas that allow the calculation of the necessary water values, with regard to the algorithms used, the whole Machine Learning solution was based on previous research on the algorithms and the development of the dataset. This may lead to a poor adaptation of the developed technology to the solution where it will be applied.

III. Dataset & Data Pre-processing

In order to develop the correct algorithm that can predict the best time of day for water administration, it is necessary to compose a correct and useful dataset.

The dataset used, as said before, was created from a range of data previously collected through the implementation of the system studied in [2]. Furthermore, the data were complemented by values provided by the Portuguese Sea and Atmosphere Institute (IPMA).

This data contains a vast number of features, as can be seen in Table I. All entries were individually analyzed, and extra features were added to each one, making the dataset richer and more substantiated in order to facilitate the process of training the algorithms under study and achieving better results.

Within these are the features *"Is_favorable"*, which classifies the ground conditions at that specific time as favorable or unfavorable for sustainable irrigation, *"Need_Irrigation"* which indicates if irrigation is needed and *"Had_Irrigation"* where it is indicated if the land has already been watered. These values were calculated based on the sensor data collected. Finally each entry was manually classified with the best irrigation hour according to the real time sensor data, within the label *"Suggested_Hour"*.

After conclusion, the dataset used has 105217 entries.

TABLE I
Dataset properties

Feature	Description
Year	Timestamp Year
Month	Timestamp Month
Day	Timestamp Day
Hour	Timestamp Hour
Temperature	Air Temperature [ºC]
Relative_Humidity	Air Humidity [%]
Total_Precipitation_Low	Precipitation [mm/H]
Wind_Speed	Wind Speed [km/h]
Wind_Direction	Wind Direction [º]
Soil_Humidity	Soil Moisture [%]
Had_irrigation	Field irrigated [0/1]
Need_Irrigation	Field needs irrigation [0/1]
Is_Favorable	Conditions favorable for irrigation [0/1]
Suggested_Hour	Suggested irrigation hour

IV. Machine Learning Classification Algorithms

Classification algorithms are the chosen ones to use in this study. Classification is the process of predicting decision values in the qualitative or category class of a given data point.

Random Forest (RF) is a tree-based method that conglomerates several self-determining decision trees developed for classification and regression. Through the combination of the various trees it is able to understand which is the best option, being the main objective to reach one in pure i.e, a node formed by a single class, giving it high predictive capabilities [7].

Decision Trees (DT) are tree based methods in which each path begin in a root node representing a sequence of data divisions until reach a Boolean outcome at a leaf node. These methods can be applied for classification and regression. The final goal of this method is to reach a model that can predict the search value for that specific scenario by learning simple decision rules [8].

Support Vector Machines (SVM) are a set of supervised learning methods developed for classification, regression and outlier's detection which is known by his high effective in high dimension spaces and for its use for training points in the decision function, being also memory efficient [9].

For the study of Neural Networks algorithms, which are defined as computational models of nervous system, the Multilayer Perceptron (MLP) method was used, which is a supervised learning method that learns a function $f(.) : R^m \rightarrow R^o$ by training on a data set, where m is the number of dimensions to input and o the number of dimensions for output. These MLP networks are characterized by being general-purpose, flexible and non-linear. Their complexity can be changed according to their application by varying the number of layers and units of each layer [10], [11].

Regarding XGBoost, this is a boosting tree based method which in turn are based on decision trees. Considering that

the linear combination for multiple trees capabilities that can well fit the training data and describe the complex non linear relationship between input and output data, makes this method considered one of the best methods in statistical learning [12].

V. Methodology

The methodology followed in the development and improvement of the algorithms previously described was divided into 4 phases.

In a first phase the dataset was built, following the steps previously described. After the dataset was completed, all values were analyzed and each entry classified as favorable or unfavorable for water administration, whether or not it needs water, if the land has already been watered and the best hour to irrigate (*"Is_favorable", "Need_Irrigation", "Had_Irrigation" and "Suggested_Hour"* respectively).

In a second phase, with the completed dataset, and before testing the various algorithms, a test was made on the importance of each feature of the dataset, allowing to understand which are the most important and which should not be considered at the time of training, in order to optimize the dataset and leading to the elimination of noise.

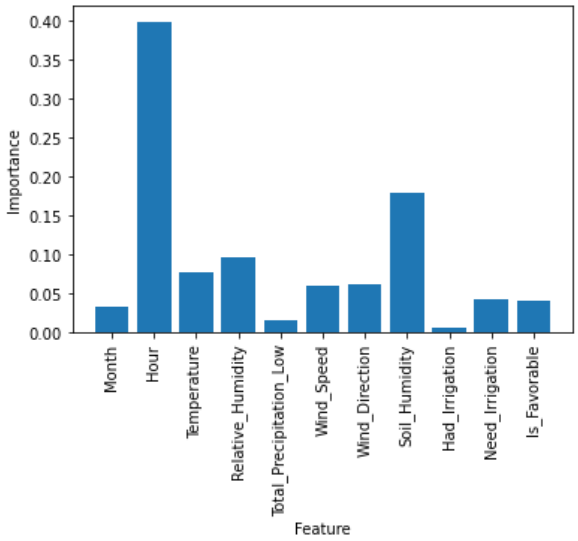

Fig. 1. Feature importance

As can be seen in Figure 1, the features *"Day", "Need_Irrigation", "Month", "Total_Precipitation_Low" and "Had_Irrigation"* have a low importance for the training and later result of the algorithms, so they were discarded. This discarding also results in a shorter time in the training of the algorithms and no significant variation was observed in the results obtained after the discarding.

The third phase of the methodology consisted in training the five algorithms under study using the parameters of each one with default values. In order to study and improve these algorithms, scikit-learn was used. This is an open source Machine Learning library developed for Python implementation [13]. For the implementation of XGBoost, a library made available by this same algorithm was used. It implements machine learning algorithms under the Gradient Boosting framework [14]. In this phase the goal is to understand which algorithms have better results using the dataset to predict the best irrigation time.

In the fourth and last phase, the best algorithms from the previous phase were exhaustively tested in order to understand what would achieve better accuracy values. In these tests an hyperparametrization tuning was done to each algorithm, in order to understand the best scenario possible. For this, a method provided by scikit-learn called RandomizedSearchCV was used, which performs the fit and training of the algorithm under study, calculating which parameters are best suited to it [15].

The final algorithm is then adapted to be running on the central server of the system under test, where the database is hosted and where the values will be received in real time. Thus, the algorithm receives the collected values and calculates the best time for water administration. When crossing the resulting values of the algorithm developed with the algorithms previously spoken and already developed by [2] that calculate the amount needed to administer the terrain under analysis, it will lead to a better management an to higher water saving values, being this the main objective.

Figure 2 shows the flowchart of the described methodology.

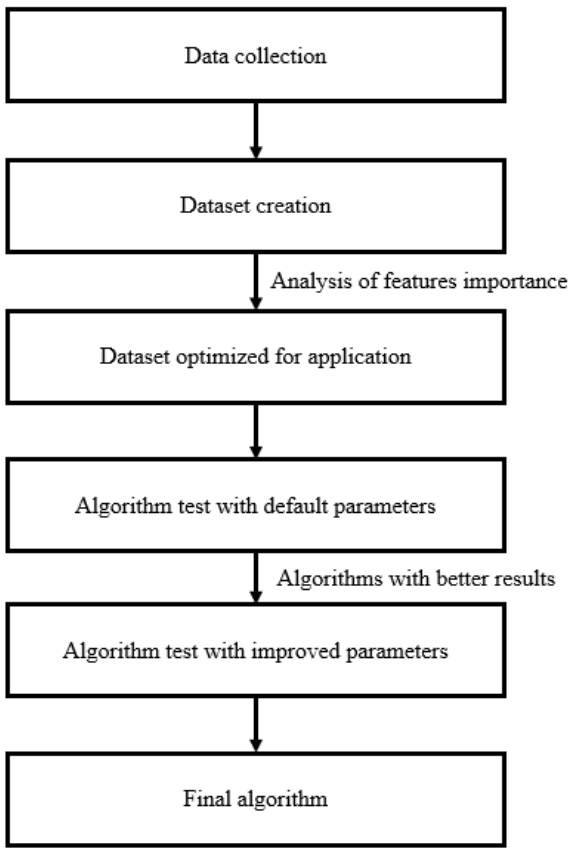

Fig. 2. Methodology Flowchart

VI. Results & Discussions

In order to put into practice the described methodology, it is necessary to precede the training and subsequent testing of the various algorithms under study to understand what best suits the intended application.

To check if the algorithms have a good applicability when receiving real values, the dataset was divided in two parts. One with 70% of the dataset, that will be used to train the algorithm, and other with the remaining 30%, will be used to test the accuracy of the trained models.

Table II shows the default parameters used for each of algorithms.

The accuracy results obtained for each model can be seen in Table III. As can be seen, the accuracy values of each algorithm trained were quite varied, which allowed to see which were best suited to the dataset and application under study.

It is possible to notice that the SVM does not fit in the proposed goal, having a low accuracy value and, even with the variation of some parameters, it would not be possible to achieve acceptable values. As for the other algorithms, although higher when compared to SVM, it is possible to conclude that only two stand out even before any optimization. As such XGBoost and RF will be further evaluated, since DT and NN, even after optimized, will not be able to reach better accuracy.

Moving forward to the next training phase, using only XGBoost and RF, and based on the parameters shown in Table II and the documentation for each of these algorithms, the hyperparametrization tuning was made only for the most important parameters, mainly those who have a numerical value. Table IV shows the tuning options for each of the selected algorithms.

The results obtained with the hyperparametrization tuning, including the best parameters settings and accuracy, can be observed in Table V. Through the analysis of the results, after the optimization of the various parameters of the two algorithms under study it is then possible to conclude that by choosing the best parameters, instead of the default configuration, it is possible to improve the accuracy of the models. Although is is not a huge improvement, 1% for XGBoost and only 0.1% for RF, this improvement can lead to the saving of a huge amount of water.

In terms of the algorithm that has the best accuracy for the situation and dataset tested is the XGBoost, which will be used for the solution.

VII. Conclusion

In this article a study of machine learning algorithms was made in order to understand which will have the higher accuracy when classifying the ideal hour to irrigate an agricultural field, based on local sensor and weather data. The algorithms tested included Random Forest, Neural Network, XGBoost, Decision Trees and Support Vector Machine.

The literature on this topic showed that research is already being done to calculate the amount of water to be administered to the agricultural field, however the time of day at which this administration was done continues to be decided by the owner and in a poorly founded way.

A methodology was followed to obtain a suitable dataset for the study and several scenarios were explored in order to understand which algorithm best suited the situation under study, and it was concluded that XGBoost was the most suitable.

After the optimization of the tested algorithm it was possible to reach an accuracy in the order of 87%, which leads to believe that the final result can improve water management and consequent savings of this natural resource.

Comparing to the results obtained by [6], with 80% accuracy using RF, our methodology obtains better results with XGBoost. Also, in terms of comparison, when using RF, our methodology also gets better results, with 84% accuracy.

As future work for this study is included the implementation of the algorithm developed in a real situation in order to test the water saving values that can be achieved and also the attempt to optimize even further the system. In order to be even more effective, and since the proposed algorithm predicts the best hour of the day for irrigation, the developed algorithm should be implemented in parallel with algorithms that calculates the amount of water needed to manage the land under study. All of these implementations should have in mind the collection of data in real time, leading to a quick response for any type of situations.

References

[1] Food and Agriculture Organization of the United Nations, "Water for Sustainable Food and Agriculture," 2019. [Online] Available: http://www.fao.org/3/a-i7959e.pdf, (visited 20/08/2020).

[2] J. Cardoso, A. Glória, and P. Sebastião, "A Methodology for Sustainable Farming Irrigation using WSN, NB-IoT and Machine Learning," in *5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA20)*, 2020.

[3] A. Glória, P. Sebastião, C. Dionísio, G. Simões, and J. Cardoso, "Water management for sustainable irrigation systems using internet-of-things," *Sensors (Switzerland)*, vol. 20, 3 2020.

[4] B. Suntaranont, S. Aramkul, M. Kaewmoracharoen, and P. Champrasert, "Water irrigation decision support system for practicalweir adjustment using artificial intelligence and machine learning techniques," *Sustainability (Switzerland)*, vol. 12, 3 2020.

[5] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar, "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance," *Computers and Operations Research*, vol. 119, 7 2020.

[6] A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, and A. Sharma, "IoT and Machine Learning Approaches for Automation of Farm Irrigation System," in *Procedia Computer Science*, vol. 167, pp. 1250–1257, Elsevier B.V., 2020.

[7] M. Mamdouh, M. A. Elrukhsi, and A. Khattab, "Securing the Internet of Things and Wireless Sensor Networks via Machine Learning: A Survey," in *2018 International Conference on Computer and Applications, ICCA 2018*, 2018.

[8] F. J. Yang, "An extended idea about decision trees," in *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, pp. 349–354, Institute of Electrical and Electronics Engineers Inc., 12 2019.

[9] Q. Tang, X. Ge, and Y. C. Liu, "Performance analysis of two different SVM-based field-oriented control schemes for eight-switch three-phase inverter-fed induction motor drives," in *2016 IEEE 8th International Power Electronics and Motion Control Conference, IPEMC-ECCE Asia 2016*, 2016.

TABLE II
DEFAULT CONFIGURATION PARAMETERS

Algorithm	Parameter Configuration
XGBoost	$eta = 0.3, gamma = 0, max\_depth = 6, min\_child\_weight = 1, max\_delta\_step = 0, subsample = 1,$ $sampling\_method = uniform, colsample\_by* = 1, lambda = 1, alpha = 0, tree\_method = auto, sketch\_eps = 0.03,$ $scale\_pos\_weight = 1, updater = grow\_colmaker, refresh\_leaf = 1, process\_type = default, grow\_policy = depthwise,$ $max\_leaves = 0, max\_bin = 256, predictor =' auto', num\_parallel_tree = 1$
Random Forest	$n\_estimators = 10, criterion =' gini', max\_depth = None, min\_samples\_split = 2, min\_samples\_leaf = 1,$ $max\_features =' auto', max\_leaf\_nodes = None, bootstrap = True, oob\_score = False, n\_jobs = 1,$ $random\_state = None, verbose = 0, min\_density = None, compute\_importances = None$
Neural Network	$hidden\_layer\_sizes = (100,), activation =' relu', *, solver =' adam', alpha = 0.0001, batch\_size =' auto',$ $learning\_rate =' constant', learning\_rate\_init = 0.001, power\_t = 0.5, max\_iter = 200, shuffle = True,$ $random_state = None, tol = 0.0001, verbose = False, warm_start = False, momentum = 0.9,$ $nesterovs\_momentum = True, early\_stopping = False, validation\_fraction = 0.1, beta\_1 = 0.9, beta\_2 = 0.999,$ $epsilon = 1e-08, n\_iter\_no\_change = 10, max\_fun = 15000$
SVM (linear)	$penalty =' l2', loss =' squared_hinge', *, dual = True, tol = 0.0001, C = 1.0, multi\_class =' ovr',$ $fit\_intercept = True, intercept\_scaling = 1, class\_weight = None, verbose = 0, random\_state = None, max\_iter = 1000$
Decision Tree	$criterion =' gini', splitter =' best', max\_depth = None, min\_samples\_split = 2,$ $min\_samples\_leaf = 1, min\_weight\_fraction\_leaf = 0.0, max\_features = None,$ $random\_state = None, max\_leaf\_nodes = None, min\_impurity\_decrease = 0.0, min\_impurity\_split = None,$ $class\_weight = None, presort =' deprecated', ccp\_alpha = 0.0$

TABLE III
DEFAULT CLASSIFICATION RESULTS

Algorithm	Accuracy
XGBoost	86.57%
Random Forest	84.77%
Neural Network	81.71%
SVM (linear)	31.38%
Decision Tree	79.92%

TABLE IV
DEFAULT CONFIGURATION PARAMETERS

Algorithm	Parameter Configuration
XGBoost	$max\_depth = randint(1, 500),$ $n_estimators = randint(1, 500),$ $subsample = [0.1, 0.2, 0.3, 0.4, 0.5,$ $0.6, 0.7, 0.8, 0.9, 1],$ $tree\_method = ['auto','exact','approx']$
Random Forest	$n\_estimators = randint(1, 500),$ $criterion = ['gini','entropy'],$ $max\_depth = randint(1, 500),$ $min\_samples\_split = randint(1, 100),$ $max\_features = ['auto','sqrt','log2'],$ $bootstrap = [True, False]$

TABLE V
OPTIMIZED CONFIGURATION PARAMETERS & RESULTS

Algorithm	Parameter Configuration	Accuracy
XGBoost	$max\_depth = 48,$ $n_estimators = 324,$ $subsample = 0.5, tre\_method = auto$	87.73%
Random Forest	$n\_estimators = 212,$ $criterion = gini, max\_depth = 196,$ $min\_samples\_split = 10,$ $max\_features =' sqrt',$ $bootstrap = True$	84.74%

[10] R. Saravanan and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," in *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018*, 2019.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, 2009.

[12] G. Seif, "A Beginner's guide to XGBoost," 2019. [Online] Available: https://towardsdatascience.com/a-beginners-guide-to-xgboost-87f5d4c30ed7, (visited 27/08/2020).

[13] scikit-learn, "scikit-learn," 2019. [Online] Available: https://scikit-learn.org/stable/, (visited 27/08/2020).

[14] S. Lu, B. Wang, H. Wang, and Q. Hong, "A hybrid collaborative filtering algorithm based on KNN and gradient boosting," in *The 13th International Conference on Computer Science & Education (ICCSE 2018)*, 2018.

[15] scikit-learn, "RandomizedSearchCV," 2019. [Online] Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html, (visited 27/08/2020).