

Sustainable Development Goals: Approach Using Machine Learning and Earth Observation Data

Bruno Ferreira^a, Muriel Iten^b, Rui Silva^c

^abaferreira@isq.pt

^bmciten@isq.pt

^crgasilva@gmail.com

RESUMO: O objetivo deste trabalho é avaliar, compreender e alcançar os Objetivos de Desenvolvimento Sustentável (ODS) através de dados de Observação da Terra (OT) e técnicas de Machine Learning (ML). Para o caso de estudo selecionado, os parâmetros analisados foram: diversos índices de vegetação e os valores das bandas espectrais, que foram extraídos de dados OT (Sentinel-2) e validados com diferentes abordagens de ML. Os resultados obtidos nas diferentes abordagens sugerem que a melhor técnica de classificação, bem como de regressão corresponde à Fusão de técnicas. No geral, constata-se que a OT desempenha um papel fundamental na monitorização e execução dos ODS, devido à sua relação custo-eficácia, à riqueza da informação, e o sucesso do ML na análise dos dados. A aplicabilidade das técnicas de ML aliadas aos dados de OT provou, com o caso de estudo, que estas podem contribuir para os ODS, assim como para outros fins.

PALAVRAS-CHAVE: Objetivos De Desenvolvimento Sustentável; Observação Da Terra; Dados de Observação da Terra; Machine Learning; Agricultura.

ABSTRACT: The main objective of this work is to evaluate and understand as well as to achieve the Sustainable Development Goals (SDGs) through Earth Observation (EO) data and Machine Learning (ML) techniques. For the selected case study, the parameters analyzed were: vegetation indices and the spectral bands' values, which were extracted from EO data (Sentinel-2) and validated with different ML approaches. The results obtained in the different ML approaches suggest that the best classification technique, as well as the best regression technique corresponds to the fusion of techniques. Overall, it is observed that EO plays a key role in monitoring and executing the SDGs, due to its cost-effectiveness, the wealth of information and the success of the ML in data analysis. The applicability of ML techniques combined with EO data proved, within the case study, that these can contribute to the SDGs and can be used for other purposes.

KEYWORDS: Sustainable Development Goals; Earth Observation; Earth Observation Data; Machine Learning; Agriculture.

1. Introduction

The concept of Sustainable Development (SD) was developed in the 1960s when it became evident that environmental problems could be caused by economic and industrial development. In 1972, a first report was published and presented at the UN, which aimed to increase public understanding of contemporary problems by addressing the impact of the 5 main trends of global concern (i.e., accelerated industrialization, rapid population growth, widespread malnutrition, resource depletion non-renewable resources and a deteriorating environment), consequences of the world's rapid development [1]. This report, called the Meadows Report, [1] was heavily criticized at the time for advocating non-growth for developing countries.

[2]. Later, in 1987, the Brundtland Report (BR)[3] defined the SD concept as development that meets the essential needs of the present without compromising future generations from meeting their own needs. In 2000, the Millennium Development Goals (MDGs) set 8 goals to fight poverty and hunger, achieve gender equality and improve the health sector. [4]. By 2015, the MDGs [5] have driven SD progress, including improvements in health and education services, reduced hunger and equity gaps, and higher levels of coverage in high-investment interventions [6], [7]. However, it remained incomplete, and in 2012 new goals were established, called the Sustainable Development Goals (SDGs) [5], setting 17 unique goals, representing an urgent call to shift the world onto a more sustainable path. [8], [9].

Earth Observation (OT) plays an important role in supporting progress on many of the SDGs [10], [11]. According to the United Nations [12], it is advantageous to use TO data to produce and support official statistics to complement traditional data sources. Satellite images may be one of the only cost-effective technologies capable of delivering data on a global scale. [13], [14]. The huge amount of data, provided by OT sources, needs to be analyzed and processed with appropriate methods and tools to provide robust indicators about DS. The growth of the field of Machine Learning (ML), which is present in our daily lives [15], [19], is constantly creating new opportunities to monitor and evaluate humanitarian efforts, plays an essential role in the analysis of applied satellite images to the SDGs. In fact, most of the methods used to process OT data are based on ML [10], [20] given its ability to process huge amounts of data and its unique classification and prediction features.

The work is structured as follows: chapter 1 (the current one), introduction of concepts related to the main research topics of this work, as well as its objectives. Chapter 2 describes the methods and materials used, as well as practical developments and the methodology to be applied to the case study. Then, in chapter 3, the case study is described and the results presented and discussed. Chapter 4 presents the main conclusions of this work, considering the results achieved.

1.1 Objectives

This work intends to explore new ML and OT methodologies and techniques to optimize (eg time, costs and resources) and improve monitoring processes to achieve the SDGs. This is achieved through 2 case studies, which it is necessary to map and/or forecast natural resources, acquiring available OT data and working them with ML techniques in order to support the realization and monitoring of the SDGs (2, 8, 9 e 12) [5], as well as the objectives of the European Space Strategy (maximizing the benefits of space for EU society and economy; promoting an innovative and globally competitive European space

sector; strengthening Europe's autonomy in accessing and using space in an environment safe and secure, reinforce Europe's role as a global player and promote international cooperation)[21]. The main objective of this work is achieved through different objectives:

- Review and analysis of the current state of the art in Sustainable Development Goals (SDGs), Earth Observation (OT) and Machine Learning (ML);
- Analysis of European and public OT data sources;
- Identification of the most appropriate ML techniques for the analysis and classification of OT data and images (literature survey);
- Development and evaluation of ML techniques;
- Prediction and classification of agriculture related problems (i.e. phytosanitary problems - healthy or infected crops), for sustainable and sufficient production.

1.2 Alignment with the Sustainable Development Goals

The results of this work contribute to the following SDGs:

- ODS 2 - Eradicate hunger and promote sustainable agriculture: through ML results that allow the prediction of diseases, quality and crop yields. Thus, it is possible to achieve greater efficiency of resources and production;
- ODS 8 - Promoting an inclusive and sustainable economy, employment and decent work for all: keeping jobs, as pests such as *Xylella Fastidiosa* put several thousand jobs in the area across Europe at risk;
- ODS 9 - Industrial innovation: by providing information on the state of the crop, allowing for more effective, responsible and recommended agricultural processes and procedures
- ODS 12 - Sustainable consumption and production: through information that allows the reduction of the environmental impact or the ecological footprint, such as the prediction of tree diseases, which can contribute to the reduction of chemical products used in agriculture;

1.3 Applied methodology

The theoretical basis of this research corresponds to an action-research methodology, as this allows the researcher to act on a real problem, trying to solve it, as it is studied/investigated. This aspect is crucial for this work due to the novelty of the investigated fields and problems. The research procedure, as well as the applied methodology (described in this chapter), consider five main steps (**figure 1**), to be later applied and adapted to the case study whose objective is based on mapping and predicting agricultural crops to provide support monitoring through risk maps.

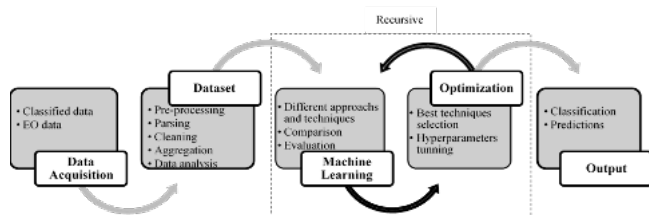


Fig. 1 – Steps of the applied methodology.

Having established the basis of the theoretical perspective (state of the art in ODS, OT techniques and ML models), the general approach presented in **figure 1** is implemented for both case studies. Firstly, data acquisition, classified data for the case study, is carried out through contacts with specialists and researchers. The remaining data (i.e. input data), necessary to build the dataset for the ML algorithms, are collected from OT data sources (European and public) suitable for the purpose of each case study. Then, to build the dataset, some tasks (i.e. processing, cleaning, filtering and aggregating data from different data sources) must be performed. Then, several ML algorithms are applied to the final dataset, where a comparison and selection of the most appropriate categories and ML techniques is performed, also considering a better understanding of which are the most relevant characteristics (those that best explain and contribute for the result/objective of each case study).

Each selected ML model must be applied 3 times, using 80% of the data for training and 20% for

testing the techniques. In order to guarantee a good representation and the same conditions for each compared ML technique, the data must be previously “mixed/shuffled”. The selection of ML categories used in this work (i.e., classification and regression techniques) is based on the type of output (i.e., categorical, continuous, or binary) and the expected result in each case study (i.e., classification or forecast).

Lastly, regarding the optimization process, all approaches are applied first without any optimization (i.e. using default parameter values). Then, the optimization of the techniques is carried out, where the parameters are adjusted to obtain better results. After this process, the final result is produced, where the techniques must be applied with their best parameters..

2. Methods and materials

A systematic review and analysis of published articles, in the English language, was conducted using ScienceDirect and Google Scholar. The survey considers research topics such as: sustainable development, sustainable development goals, Earth observation and machine learning. To explore a wide range of different applications, the name of various fields (e.g. agriculture, climate, forestry and mineral prospecting) were researched in combination with Earth observation, and compared to adding machine learning. The search was refined to support results of relevance, considering the research and most recent case studies, while maintaining historical accounts and agreements.

3. Development methodology for the case study

In order to achieve the main objective of this study and respond to the gap identified by Ferreira ^[22] (high need for new ML methods and techniques to be applied to OT data for the execution and monitoring of the SDGs), in addition to common approaches, new ML techniques are developed: (i) fusion of techniques and (ii) Random Forest based on non-binary decision

tree. To evaluate the results and performance of these techniques, the (i) common classification and (ii) regression techniques, together with the (iii) Non-Binary Decision Tree (previously developed by Ferreira [23]) are implemented.

- **Sorting Techniques:** For the evaluation of the results, 16 classification techniques from the Scikit-learn library (for example, Random Forest and K-Nearest Neighbor), were implemented and applied to the datasets. Then, a comparison and optimization of the techniques is performed, in order to choose the most suitable technique for this application.
- **Regression techniques:** This approach is the same as the previous one, but instead of using classification techniques, it uses 16 regression techniques (for example, Regression Trees and Support Vector Regression). Furthermore, data used in combination with regression techniques are normalized (i.e. Min-Max normalization [24], where attributes/variables are scaled to a range between 0 and 1.), with the aim of improving its performance [25].
- **Non-binary decision tree (NBDT):** In order to overcome the inherent limitations of ID3 decision trees [26] (manipulation of continuous variables) an adaptation of the original algorithm was developed and implemented using the same metrics, but allowing its application in data with continuous variables. Furthermore, this technique also proves to be efficient and accurate using only a small part of the number of attributes present in the original dataset [23].
- **Random Forest based on non-binary decision tree (NBRF):** This proposed algorithm was developed based on the NBDT [23], presented above, and on the original RF [27]. Thus, instead of producing a single tree, it produces n trees and, for each tree, a series of observations and attributes are randomly removed, thus producing different trees. After all trees are completed, the mode (most frequent value) is applied to the results of all trees for each observation. If two or more outputs have the same

mode value, the result from the tree with the best performance is selected.

- **Fusion Techniques:** This proposed approach is based on the Random Forest (RF) concept [27], where each tree produced contributes to the final result. Thus, in this approach, different classification or regression techniques are used with a similar concept, which allows producing more reliable results when compared to the application of a single technique, as the outlier of one technique will be minimized by the others.

4. Case study – *Xylella Fastidiosa* in olive groves

As Europe is the largest producer of olive oil in the world, where Portugal plays a key role, and the fact that the species most affected by *Xylella Fastidiosa* (Xf) is olive trees, this case study aims to predict or classify areas with low and high risk of occurrence of Xf in olive groves. The identification of Xf is essential for its eradication, however, it is quite difficult, as the visual symptoms can be confused with numerous other diseases and pests, or even with drought. [28]. Furthermore, this disease may only manifest visual symptoms years after the date of infection. The only reliable way to identify Xf is through PCR tests. Due to its rate of spread, difficult identification and the fact that there is no cure, it has become important to understand its behavior and have early warnings of areas with a higher risk of Xf. Thus, OT data have become increasingly used [28], [29], where the calculation of vegetation indices allows the production of several insights on quantitative and qualitative assessments of vegetation cover, vigor, health and growth dynamics [30].

In this case study, two types of data were needed: in-situ data: related to the presence or absence of Xf and OT data: related to the condition and health of the trees (Sentinel-2).

4.1 Results

To compare the different approaches, the True Negative accuracy metric was chosen for the classification techniques (it was considered more relevant to know exactly the areas where Xf does not occur, for this reason the general accuracy comes in the background) while which for the regression techniques was the Root Mean Squared Error.

Observing the results of the five approaches (presented in **table 1**), it is possible to infer that, among all the approaches, the best one for classification corresponds to the Fusion of classification techniques (True Negative accuracy of 100% and overall accuracy of 92.87%) and optimized Extra Tree Classifier (True Negative accuracy of 99.178% and an overall accuracy of 93.727%); and the best regression technique is the Fusion of Regression Techniques (RMSE 0.148) followed by the optimized K-Nearest Neighbor Regressor (RMSE 0.206). The NBDT and NBRF approaches were discarded after the first tests, as their results, despite being positive (especially the NBRF with 84.01% overall accuracy), fell short of the other approaches. However, it is important to highlight the fact that the proposed techniques (ie, Technique Fusion and NBRF) achieved superior performance when compared to previous techniques (ie, classification and regression techniques and, NBDT).

Tab. 1: Results of different ml approaches.

Approach	Accuracy	True Negative Accuracy	True Positive Accuracy	RMSE
Rating* (ETC)	92,74%	99,72%	43,56%	-
Optimized Rating* (ETC)	93,73%	99,18%	45,78%	-
Classification Fusion	92,87%	100%	41,11%	-
Regression* (ETR)	-	-	-	0,228
Optimized Regression* (K-NN R)	-	-	-	0,206
Regression Fusion	-	-	-	0,148
NBDT** (0.1)	56,71%	-	-	-
NBRF** (0.1)	84,01%	-	-	-

*result of the best technique.

** data resolution.

4.2 Discussion

The results presented above made it possible to predict or classify and map areas with higher and lower risk of Xf, allowing:

- Reduce monitoring and testing costs;
- Reduce the use of chemicals;
- Keep jobs;
- Increase or maintenance of production.

Data quality has a great influence on the performance of all ML techniques. Some of the in-situ data lacks accuracy in location, leading to having not a single tree but a cluster of trees, which means that the models will predict or classify a cluster rather than a single tree. In addition, there is another problem related to the time of infection of Xf, since it is not possible to establish a starting date. Symptoms present on the tree may not be visually detected for many years. In addition to these requirements, balanced data - between infected and uninfected trees - is also needed to train the algorithms in an unbiased way.

5. Conclusions

Sustainable Development (SD) is unavoidable for the growth of any society and country, and which highlights the fact that the Sustainable Development Goals (SDGs) are crucial for the future of our planet. This work highlights the importance of achieving DS through Earth Observation (OT) data and Machine Learning (ML) techniques. The monitoring and evaluation aspects related to SD leveraged the OT data collection methods. OT is a cost-effective technology capable of providing data at a global level and thus enabling a different perspective on the SDGs. Due to the high amount of OT data, its analysis becomes more powerful and possible when combined with ML techniques. For this work, two case studies were used to demonstrate the practical benefits of using ML techniques on OT data to achieve the SDGs.

The results obtained suggest that, in the case study (with the objective of mapping and predicting the risk or occurrence of *Xylella Fastidiosa* in olive groves), of all the techniques applied, the ones that proved to be

the most appropriate were the Fusion of classification techniques and the Extra Tree Classifier. With regard to the best regression technique, the Fusion of regression techniques was identified. These results can leverage and optimize processes in these environments and industries, thus contributing to the monitoring and achievement of SDGs.

In summary, it is possible to state that the innovative methodology presented in this study - combination of OT data and ML techniques (in particular the common techniques of classification and regression, and the fusion technique for classification and regression),

successfully allow , contribute to and monitor the SDGs (2, 8, 9 and 12). On the one hand, OT data plays a key role in translating real-world problems into variables and data that can be used by ML techniques. Furthermore, the proposed ML approaches or techniques (i.e. Random Forest based on non-binary decision Tree, and merging of the techniques) performed better when compared to the generic/original ones (i.e. decision Tree not binary, and common classification and regression techniques. Finally, the results obtained also suggest that these methodologies and techniques can be applied to other problems, environmental and even to other SDGs.

References

- [1] D. H. Meadows, D. L. Meadows, J. Randers, and W. W. Behrens III, "The Limits to Growth: A Report to The Club of Rome," *Formaliz. Math.*, 1972, doi: 10.2478/v10037-009-0024-8.
- [2] M. Mesarovic and E. Pestel, "Mankind at the turning point: The second report of the club of Rome," *Econ. Anal. Policy*, 1975, doi: 10.1016/S0313-5926(76)50005-1.
- [3] WCED, "Report of the World Commission on Environment and Development: Our Common Future: Report of the World Commission on Environment and Development," 1987. doi: 10.1080/07488008808408783.
- [4] United Nations, "United Nations Millennium Declaration," *Gen. Assem.*, no. September, p. 9, 2000, [Online]. Available: http://www.un.org/en/events/pastevents/millennium_summit.shtml.
- [5] United Nations, "Transforming Our World: The 2030 Agenda for Sustainable Development," *A New Era Glob. Heal.*, no. 1, pp. 1–64, 2015, doi: 10.1891/9780826190123.ap02.
- [6] C. Moucheraud et al., "Countdown to 2015 country case studies: what have we learned about processes and progress towards MDGs 4 and 5?," *BMC Public Health*, vol. 16, no. S2, p. 794, Sep. 2016, doi: 10.1186/s12889-016-3401-6.
- [7] United Nations, *The Millennium Development Goals Report 2015*. New York: UN, 2015.
- [8] R. G. Gusmão Caiado, W. Leal Filho, O. L. G. Quelhas, D. Luiz de Mattos Nascimento, and L. V. Ávila, "A literature-based review on potentials and constraints in the implementation of the sustainable development goals," *Journal of Cleaner Production*, vol. 198. Elsevier Ltd, pp. 1276–1288, Oct. 10, 2018, doi: 10.1016/j.jclepro.2018.07.102.
- [9] United Nations Development Programme, "Background of the Sustainable Development Goals," United Nations Development Programme, 2017. <https://www.undp.org/content/undp/en/home/sustainable-development-goals/background.html> (accessed Feb. 15, 2019).
- [10] A. Andries, S. Morse, R. Murphy, J. Lynch, E. Woolliams, and J. Fonweban, "Translation of Earth observation data into sustainable development indicators: An analytical framework," *Sustain. Dev.*, vol. 27, no. 3, pp. 366–376, May 2019, doi: 10.1002/sd.1908.
- [11] GEO, *Earth Observations in support of the 2030 Agenda for Sustainable Development*, no. March. Japan Aerospace Exploration Agency, 2017.
- [12] United Nations, "Earth Observations for Official Statistics: Satellite Imagery and Geospatial Data Task Team report," no. December, p. 170, 2017, Accessed: May 20, 2019. [Online]. Available: https://unstats.un.org/bigdata/taskteams/satellite/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf.
- [13] K. Murthy, M. Shearn, B. D. Smiley, A. H. Chau, J. Levine, and D. Robinson, "SkySat-1: very high-resolution imagery from a small satellite," Oct. 2014, doi: 10.1117/12.2074163.

- [14] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Testing the Race Model Inequality in Redundant Stimuli With Variable Onset Asynchrony," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 35, no. 2, pp. 575–579, 2016, doi: 10.1037/a0013620.
- [15] A. Alghoul, S. Al Ajrami, G. Al Jarousha, G. Harb, and S. S. Abu-Naser, "Email Classification Using Artificial Neural Network," 2018. Accessed: Jul. 07, 2020. [Online]. Available: www.ijeais.org/ijaer.
- [16] D. van Thiel and W. F. F. van Raaij, "Artificial intelligence credit risk prediction: An empirical study of analytical artificial intelligence tools for credit risk prediction in a digital era," *Journal of Risk Management in Financial Institutions*, 2019. <https://www.ingentaconnect.com/content/hsp/jrmfi/2019/00000012/00000003/art00008> (accessed Jul. 07, 2020).
- [17] D. Trong, H. W. Lauw, Y. Fang, and D.-T. Le, "Modeling contemporaneous basket sequences with twin Modeling contemporaneous basket sequences with twin networks for next-item recommendation networks for next-item recommendation Part of the Artificial Intelligence and Robotics Commons, Databases and Information Systems Commons, and the E-Commerce Commons Citation Citation Modeling Contemporaneous Basket Sequences with Twin Networks for Next-Item Recommendation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Jul. 2018, vol. 3414, no. 3420, p. 3414, Accessed: Jul. 07, 2020. [Online]. Available: https://ink.library.smu.edu.sg/sis_research.
- [18] S. Kumer Roy, A. Kanti Ghosh, A. Siddika Asa, M. Palash Uddin, M. Rashedul Islam, and M. Ibn Afjal, "Bengali consonants-voice to text conversion using machine learning tool," 2017. Accessed: Jul. 07, 2020. [Online]. Available: <http://www.ijrcee.org>.
- [19] G. Lugano, "Virtual assistants and self-driving cars," Jul. 2017, doi: 10.1109/ITST.2017.7972192.
- [20] T. Landry et al., "Applying Machine Learning to Earth Observations In A Standards Based Workflow," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2019, pp. 5567–5570, doi: 10.1109/IGARSS.2019.8898032.
- [21] European Commission, "Space Strategy for Europe," Brussels, 2016.
- [22] B. Ferreira, M. Iten, and R. G. Silva, "Monitoring sustainable development by means of earth observation data and machine learning: a review," *Environ. Sci. Eur.*, vol. 32, no. 1, p. 120, Dec. 2020, doi: 10.1186/s12302-020-00397-4.
- [23] B. Ferreira, R. G. Silva, and V. Pereira, "Feature selection using non-binary decision trees applied to condition monitoring," *IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA*, vol. 1, pp. 1–7, 2017, doi: 10.1109/ETFA.2017.8247642.
- [24] L. Al Shalabi, Z. Shaaban, and B. Kasasbeh, "Data Mining: A Preprocessing Engine," *J. Comput. Sci.*, vol. 2, no. 9, pp. 735–739, Sep. 2006, doi: 10.3844/jcsp.2006.735.739.
- [25] C. Chen, S. A. Vannoy, A. Adeyemo, and H. Wimmer, "Effects of Normalization Techniques on Logistic Regression in Data Science," *J. Inf. Syst. Appl. Res.*, vol. 12, no. 2, p. 37, Aug. 2019, Accessed: Jan. 03, 2022. [Online]. Available: <http://conisar.org>.
- [26] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1023/A:1022643204877.
- [27] L. Breiman, "Random Forests," 2001, Accessed: Oct. 21, 2017. [Online]. Available: <https://www.stat.berkeley.edu/%7B~%7Dbreiman/randomforest2001.pdf>.
- [28] P. J. Zarco-Tejada et al., "Previsual symptoms of Xylella fastidiosa infection revealed in spectral plant-trait alterations," *Nat. Plants*, vol. 4, no. 7, pp. 432–439, Jul. 2018, doi: 10.1038/s41477-018-0189-7.
- [29] A. Hornero et al., "Using sentinel-2 imagery to track changes produced by xylella fastidiosa in olive trees," pp. 9060–9062, 2018.
- [30] J. Xue and B. Su, "Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications," *J. Sensors*, vol. 2017, pp. 1–17, 2017, doi: 10.1155/2017/1353691.