Evaluating the use of machine learning algorithms in environmental sensing for energy saving

Giovanni Delnevo, Gianni Tumedei, Vittorio Ghini, Catia Prandi
{giovanni.delnevo2,gianni.tumedei2,vittorio.ghini,catia.prandi2}@unibo.it
Department of Computer Science and Engineering
University of Bologna
Cesena, Italy

ABSTRACT

Coastal lagoons are complex ecosystems characterized by the interaction of several actors, that can have a significant impact on them. The SMARTLAGOON project has the primary aim of integrating novel artificial intelligence-based technologies with an efficient Internet of Things (IoT) sensing infrastructure in the Mar Menor coastal lagoon. This paper presents an approach to predict some variables (chlorophyll and turbidity) usually sensed by the smart bouy in future instants of time. Results show that machine learning algorithms can accurately predict them.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Computer systems organization** → *Sensor networks*; *Embedded systems*.

KEYWORDS

environmental sensing, machine learning algorithms, coastal lagoons, digital twin

ACM Reference Format:
Giovanni Delnevo, Gianni Tumedei, Vittorio Ghini, Catia Prandi. 2023. Evaluating the use of machine learning algorithms in environmental sensing for energy saving. In *The 2nd Workshop on Networked Sensing Systems for a Sustainable Society (NET4us '23), October 6, 2023, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3615991.3616402

1 INTRODUCTION

Coastal lagoons are highly productive ecosystems with a number of different uses and services. Making up 13% of the world's coastline and approximately 5.3% of Europe [19], they are exploited for fishing, aquaculture, saltworks, and recreational activities, and play a crucial role in retaining and purifying pollutants, which is essential for the ecology of coastal areas [13]. Being affected by hydrological, hydrodynamic, ecological, and socioeconomic processes [6], these highly complex systems are particularly susceptible to both climate-related and human-induced pressures [1].

The Mar Menor coastal lagoon, situated in southeastern Spain, is the largest saltwater lagoon in Europe and is located in a region characterized by aridity and water scarcity [7]. The basin that drains into the Mar Menor experiences extensive land irrigation through water diversion from the near Tagus River, as well as excessive use of aquifers during drought periods [20]. Additionally, the Mar Menor attracts a significant population around it, as it relies on fishing and agriculture, as well as tourism during the summer season, as key economic activities [4, 9]. The rapid economic, social, and urban changes over the past several decades, along with historical mining impacts, have had numerous adverse effects on the ecological status of the Mar Menor, raising the need for urgent protective measures [10].

In this context, the SMARTLAGOON project, funded in a H2020 call, was developed with the primary aim of integrating novel artificial intelligence-based technologies with an efficient Internet of Things (IoT) sensing infrastructure. These technologies would gather input data for innovative socio-environmental dynamic models, enabling the forecasting of both short and long-term changes in the lagoon's conditions. This information would then provide added value for management decisions aimed at safeguarding the ecosystem and services offered by the Mar Menor lagoon [2]. Not only that but the acquired knowledge can be applied to the realization of similar systems that help lagoon ecosystems all around the world.

In this paper, we describe our approach to predict chlorophyll and turbidity using the data collected by the smart bouy, taking advantage of machine learning algorithms. Being able to predict future values of the sensed variables, would allow to turn off the smart bouy at regular time intervals, reducing its operating time to better manage the battery consumption.

NET4us '23, October 6, 2023, Madrid, Spain

ACM ISBN 979-8-4007-0365-2/23/10.
https://doi.org/10.1145/3615991.3616402

Several machine learning algorithms have been evaluated in the experiments.

The remainder of the paper is structured as follows. Section 2 details the dataset and the methodology that drove this study. Section 3 presents the accuracy of the machine learning algorithms. Finally, Section 4 concludes the paper, highlighting some final remarks and future works.

2 METHODS

This Section details the dataset used, the followed methodology, and the machine learning algorithms employed in the experiments.

2.1 Dataset Description

The dataset is collected using the smart bouy deploying within the Smart Lagoon project.

With the only exception of chlorophyll (Mean_Chl_ugl) and turbidity (Mean_Turb_NTU), which are available hourly, the other variables are sensed every five minutes. Generally, they are then aggregated computing the average values. For some of them, it is available also the standard deviation or the maximum value. The sensed variables are the air temperature (Air_Temp_HS_Avg), the relative humidity (RelHumidity_Avg), the steam pressure (Vapor_Pressure_Avg), the wind speed (WS_ms_Avg, WS_ms_Std, WS_ms_Max, and WS_ms_TMx), the wind power (WS3_Avg), water temperature measured by a thermistor at different depths (0.5m - ThermTemp1_Avg, 1.5m, 2m, 2.5m, 4m, and 5m ThermTemp6_Avg), water temperature measured by oximeter at different depths (1m - Wtemp_C1_Avg, 3m, and 6.5m - Wtemp_C3_Avg), water temperature measured by conductimeter at different depths (1m - SDI_Temp_1m, 3m, 6.5m - SDI_Temp_6m), oxygen saturation at different depths (1m - O2_sat1_Avg, 3m, and 6.5m - O2_sat3_Avg), oxygen concentration at different depths (1m - O2_conc1_Avg, 3m, and 6.5m - O2_conc3_Avg), conductivity at different depths (1m -SDI_Cond_1m, 3m, and 6.5m - SDI_Cond_6m), and conductivity corrected with the temperature at 25° (1m - SDI_TempCorrCond_1m, 3m, and 6m - SDI_TempCorrCond_6m). The overall dataset is composed of 4,653 rows, recorded from the 28th of August 2022 to the 10th of April 2023.

2.2 Methodology

The pre-processing activities simply consisted in the detection of missing values and outliers. Some variables, containing many missing values, were dropped.

Initially, data analysis has been carried out to understand possible correlations between the sensed variables. This activity is particularly relevant in this context, given that the same variables (e.g., water temperature, oxygen saturation, and conductivity) are surveyed at different depths. The results of such an activity are described in the next Section.

Once the correlated variables are dropped to avoid weighting them more than once, the dataset for training activities has been prepared. The objective of this study is to be able to predict future values of chlorophyll and turbidity with the aim of reducing the amount of time that sensors have to work. Hence, they are treated as the target variables. The training examples are created using the values sensed at a given time *t*, which are used as the input data, the values of chlorophyll and turbidity at a given time *t + offset*, that are used as the output data. The higher the value of the *offset* is, the higher the energy saving is. The number of samples available for training is equal to *4,653 - offset*.

Once the dataset has been defined, it was divided into two different parts. The former one, used for the training phase, is composed of the 80% of samples while the latter one for the testing phase, is composed of the remaining 20% of subjects. No validation set has been defined since the k-fold cross-validation is used [5, 12]. Such a technique is used to reduce the bias derived from random sampling and to better tune the hyper-parameter of the various algorithms. The number of folds used in the experiments was set to ten, which is a commonly used value in the scientific literature. At each iteration of the cross-validation, data are scaled, subtracting the average value and dividing by the standard deviation. Finally, once the most promising algorithms have been identified, they are used to evaluate the test set.

With regard to the evaluation metrics used to assess the performance of the different algorithms, two metrics were selected and used: Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC). The two metrics serve different purposes. The MAE, that is the average deviation between the real chlorophyll and turbidity values and the predicted ones, is used to measure the accuracy of the predictions. Instead, the PCC is used to understand if the algorithms really learned something. In fact, a naive regressor that always returns the mean value could be able to achieve good MAE scores. Hence, since the PCC quantifies the degree of the linear association between real and predicted chlorophyll and turbidity values, is able to highlight such a situation, since its value would be low in this case.

2.3 Machine Learning Algorithms

Since chlorophyll and turbidity are continuous values, the problem was modeled as a regression one. Several algorithms were evaluated to determine which one is best suited for this case study.

In order to have a baseline comparison, Linear Regression (LR) [22] and other variations such as Lasso [18] and Elastic

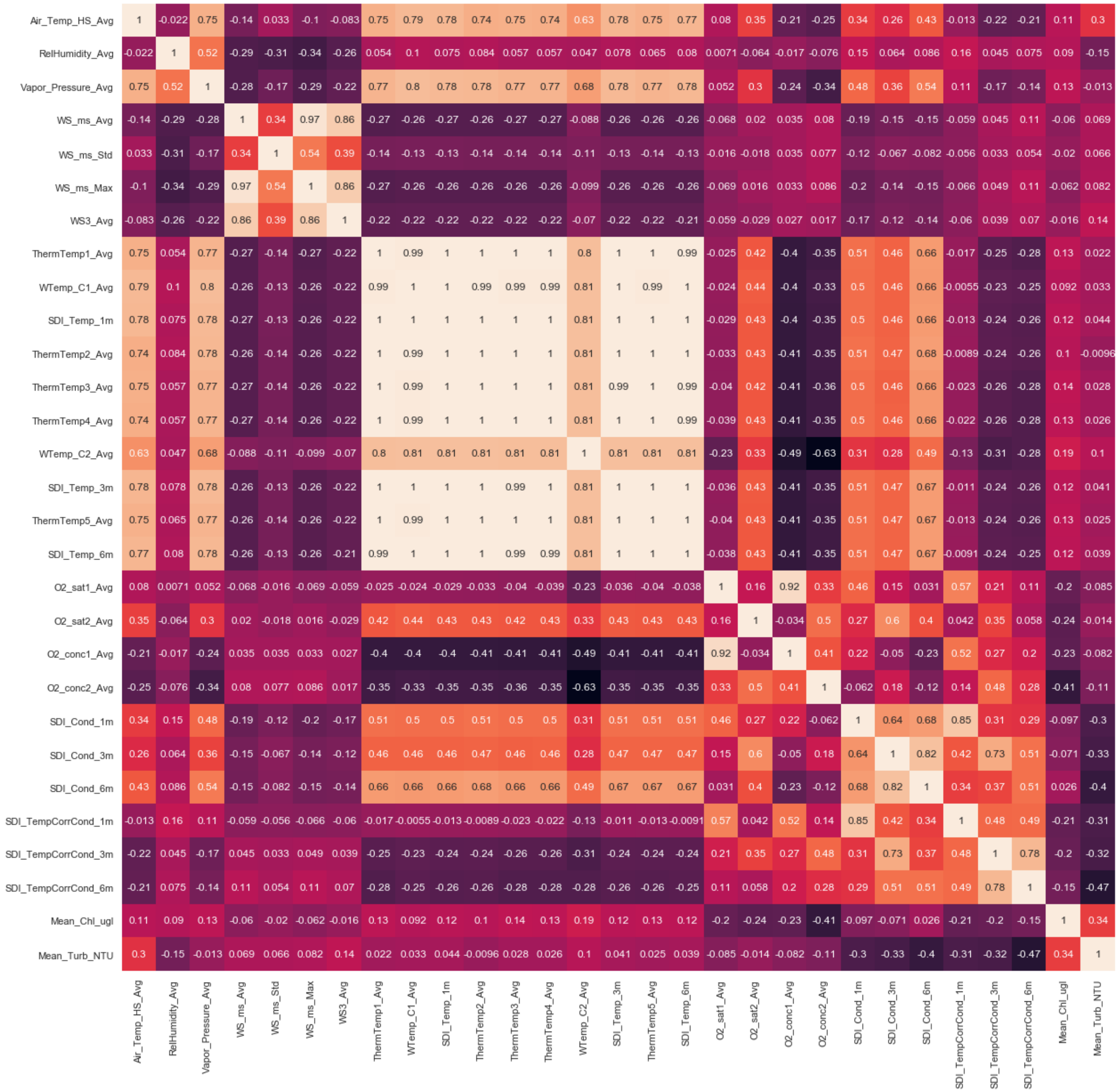

Figure 1: Correlation Analysis

Net (EN) Regression [23] were first evaluated. Then, also K-Nearest Neighbor (KNN) [8], Support Vector Machine (SVM) [14], and Multi-Layer Perceptron (MLP) [16] were analyzed. Finally, Classification And Regression Tree (CART) [3] and other ensemble algorithms, including Ada Boosting with decision trees (AB) [17], Gradient Boosting (GB) [11], Random Forest (RF) [15], and Extra Tree (ET) [21] were taken in consideration.

In the experiments, the implementation of the algorithms available in the Scikit-learn library was employed. Each algorithm has been used with its default parameters with the only exception of the random state, which was set to guarantee the reproducibility of the obtained results.

3 RESULTS

This Section presents the output of the analysis of the variables in the dataset and the accuracy of the algorithms in predicting chlorophyll and turbidity.

3.1 Data Analysis

To evaluate the correlation between the variables present in the dataset, the Pearson correlation was employed. The correlation matrix is shown in Figure 1. The colors of the matrix range from black, which indicates a negative correlation (-1) to yellow, which indicates a positive correlation (+1), passing from purple, which indicates an absence of correlation (0).

As shown, several variables are highly correlated (with the PCC higher than 0.9). Such correlations are important not only to understand which variables have to be dropped but also to highlight which sensors could not be included in further versions of the smart bouy since their sensed variables can be obtained using a simple correlation. The highly correlated variables are (in bold there is the variable that was kept for the training activities):

- The average wind speed (**WS_ms_Avg**) and the maximum wind speed (WS_ms_Max).
- All the water temperature measured by a thermistor at different depths (**ThermTemp1_Avg**, ThermTemp2_Avg, ThermTemp3_Avg, ThermTemp4_Avg, ThermTemp5_Avg), the water temperature measured by oximeter at 1m (WTemp_C1_Avg), water temperature measured by conductimeter at different depths (SDI_Temp_1m, SDI_Temp_3m, and SDI_Temp_6m).
- The oxygen concentration at 1m (**O2_conc1_Avg**) and the oxygen saturation at 1m (O2_sat1_Avg).

It is also interesting to notice that some variables measured at different depths are not correlated, such as the oxygen concentration, the oxygen saturation, conductivity, and conductivity corrected with the temperature at 25°.

3.2 Chlorophyll Prediction

First, we investigate the prediction of chlorophyll using the other variables sensed by the smart bouy, employing all the algorithms described in the previous Section. The results obtained during the cross-validation are depicted in Figure 2. As shown, LR works significantly better than LASSO and EN and has slightly worse performance than AB. Anyway, it has performance comparable with the ones obtained by the other algorithms. The best performance is achieved by ET and GB with an MAE equal to 0.11 while the other algorithms obtained an MAE in the range of 0.12 to 0.18. About the PCC scores, all the algorithms highlighted strong correlations with values in the range of 0.79 - 0.98, with the only exception of LASSO, which achieved a PCC of 0.53.

Once identified the most performing algorithms, we evaluated their performance on the test sets, without a further tuning phase. Results are reported in Table 1. As shown, in both cases, the algorithms achieved MAE scores similar to the ones obtained during the cross-validation, which highlights the ability of such algorithms to successfully predict the chlorophyll.

Table 1: Results of ET and GB during the cross-validation and on the test set for the prediction of chlorophyll.

Algorithm	Cross-validation		Test Set	
	MAE	PCCC	MAE	PCCC
ET	0,11 ± 0,01	0,98	0.126	0.883
GB	0,11 ± 0,01	0,98	0.130	0.880

3.3 Turbidity Prediction

Finally, we investigate the prediction of turbidity, with the same approach used for the chlorophyll prediction. The results obtained during the cross-validation are depicted in Figure 3. Also, in this case, LR, LASSO, EM, and AB have the worst performance while ET and GB, together with RF, got the best ones, with the MAE of 0.06 (ET) and 0.07 (GB and RF). With regard to the PCC scores, all the algorithms highlighted strong correlations with values in the range of 0.84 - 0.93, except for the ones with lower performance.

Also for the prediction of turbidity, the most performing algorithms have been evaluated on the test sets, always without a further tuning phase. Table 2 reports the results. Also in this case, even on the test set the three algorithms have similar performance with respect to the one obtained during the cross-validation, thus indicating the possibility of predicting future values of turbidity.

Table 2: Results of RF, GB, and ET during the cross-validation and on the test set for the prediction of turbidity.

Algorithm	Cross-validation		Test Set	
	MAE	PCCC	MAE	PCCC
RF	0,07 ± 0,01	0,92	0.078	0.728
GB	0,07 ± 0,01	0,92	0.08	0.726
ET	0,06 ± 0,01	0,93	0.079	0.726

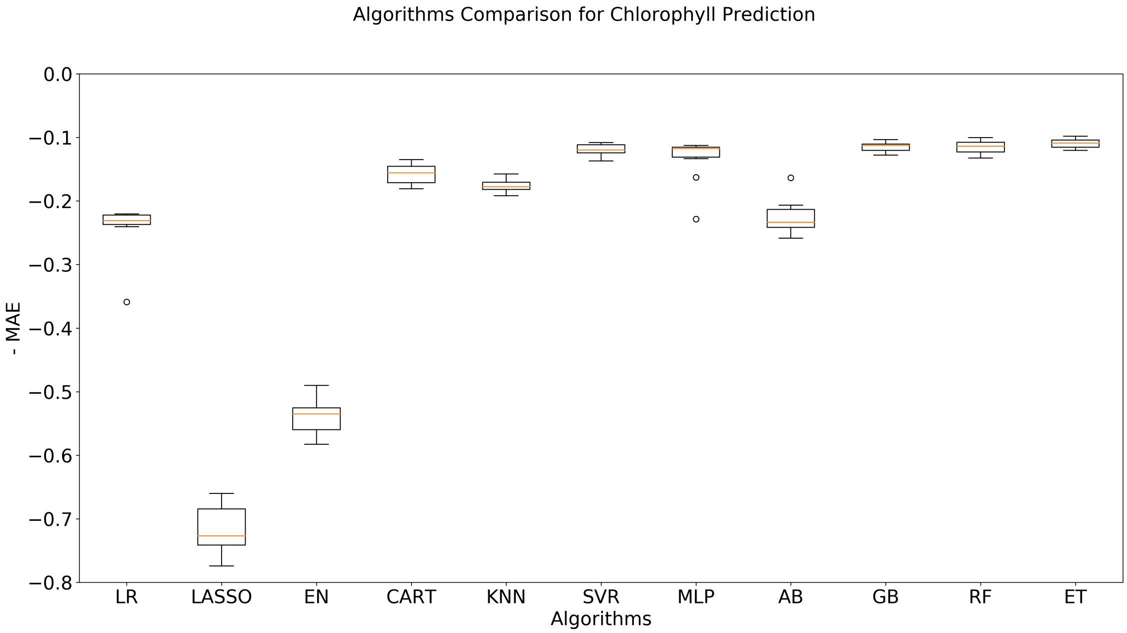

Figure 2: Comparison among algorithms performances for the chlorophyll prediction

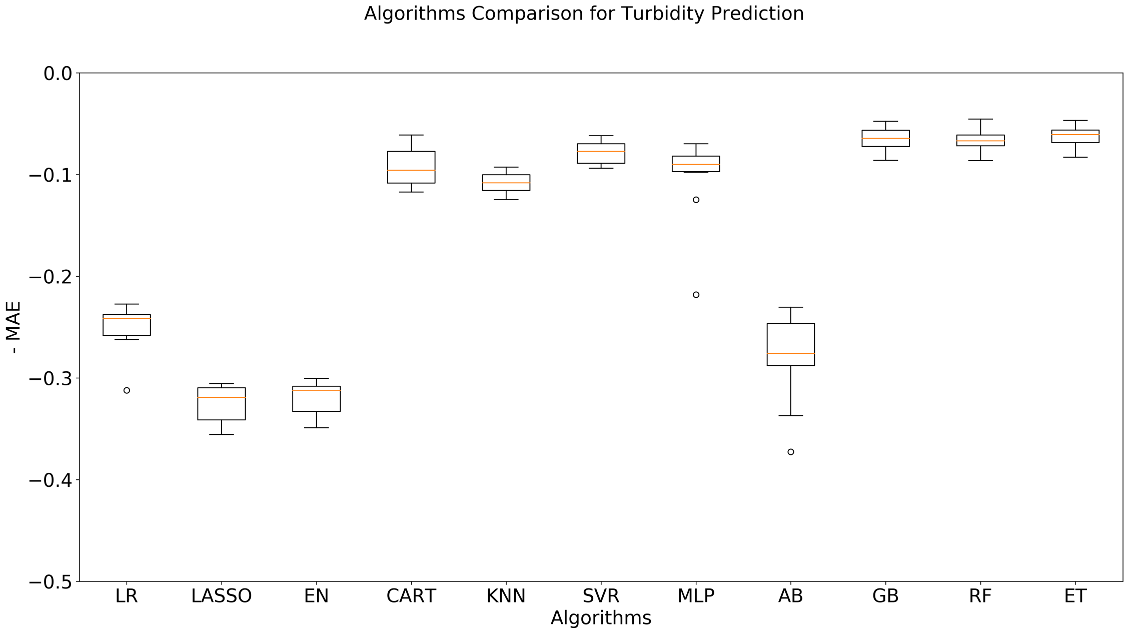

Figure 3: Comparison among algorithms performances for the turbidity prediction

4 CONCLUSION AND FUTURE WORKS

This paper presents some experiments about the prediction of some variables (i.e., chlorophyll and turbidity) sensed by the smart bouy. The results indicate that machine learning algorithms can be effectively used to predict the variable of interest. This could allow us to avoid continuously sensing all the variables, saving energy, which is a crucial aspect in IoT projects.

There are plenty of future works. First, a tuning phase of the hyper-parameters could be carried out to try to improve the presented results. Then, higher values of the *offset* could be tested to evaluate if it is possible to predict values more distant in time. Finally, a study on the prediction of all the sensed variables could be carried out. This would allow us to completely turn off the smart bouy in some fixed intervals between the different measurements.

ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017861.

REFERENCES

[1] Ana C. Brito, Alice Newton, Paul Tett, and Teresa F. Fernandes. 2012. How will shallow coastal lagoons respond to climate change? A modelling investigation. *Estuarine, Coastal and Shelf Science* 112 (2012), 98–104. https://doi.org/10.1016/j.ecss.2011.09.002 Assessing Ecological Quality in Estuarine and Coastal Systems - Functional Perspective.

[2] José M. Cecilia, Pietro Manzoni, Dennis Trolle, Anders Nielsen, Pablo Blanco, Catia Prandi, Salvador Peña Haro, Line Barkved, Don Pierson, and Javier Senent. 2021. SMARTLAGOON: Innovative Modelling Approaches for Predicting Socio-Environmental Evolution in Highly Anthropized Coastal Lagoons. In *Proceedings of the Conference on Information Technology for Social Good* (Roma, Italy) *(GoodIT '21)*. Association for Computing Machinery, New York, NY, USA, 204–209. https://doi.org/10.1145/3462203.3475925

[3] Wei Chen, Xiaoshen Xie, Jiale Wang, Biswajeet Pradhan, Haoyuan Hong, Dieu Tien Bui, Zhao Duan, and Jianquan Ma. 2017. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* 151 (2017), 147–160.

[4] Héctor M. Conesa and Francisco J. Jiménez-Cárceles. 2007. The Mar Menor lagoon (SE Spain): A singular natural ecosystem threatened by human activities. *Marine Pollution Bulletin* 54, 7 (2007), 839–849. https://doi.org/10.1016/j.marpolbul.2007.05.007

[5] Giovanni Delnevo, Silvia Mirri, Catia Prandi, and Pietro Manzoni. 2023. An evaluation methodology to determine the actual limitations of a TinyML-based solution. *Internet of Things* 22 (2023), 100729.

[6] S. García-Ayllón. 2017. Integrated management in coastal lagoons of highly complexity environments: Resilience comparative analysis for three case-studies. *Ocean & Coastal Management* 143 (2017), 16–25. https://doi.org/10.1016/j.ocecoaman.2016.10.007 The challenge of developing policies and management strategies under changing baselines and unbounded boundaries.

[7] J. García-Pintado, M. Martínez-Mena, G.G. Barberá, J. Albaladejo, and V.M. Castillo. 2007. Anthropogenic nutrient sources and loads from a Mediterranean catchment into a coastal lagoon: Mar Menor, Spain. *Science of The Total Environment* 373, 1 (2007), 220–239. https://doi.org/10.1016/j.scitotenv.2006.10.046

[8] Ikbal Gazalba, Nurul Gayatri Indah Reza, et al. 2017. Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification. In *2017 2nd international conferences on information technology, information systems and electrical engineering (ICITISEE)*. IEEE, 294–298.

[9] Noelia Guaita-García, Julia Martínez-Fernández, Carlos Javier Barrera-Causil, and H. Carl Fitz. 2022. Stakeholder analysis and prioritization of management measures for a sustainable development in the social-ecological system of the Mar Menor (SE, Spain). *Environmental Development* 42 (2022), 100701. https://doi.org/10.1016/j.envdev.2022.100701

[10] Patricia Jimeno-Sáez, Javier Senent-Aparicio, José M. Cecilia, and Julio Pérez-Sánchez. 2020. Using Machine-Learning Algorithms for Eutrophication Modeling: Case Study of Mar Menor Lagoon (Spain). *International Journal of Environmental Research and Public Health* 17, 4 (2020). https://doi.org/10.3390/ijerph17041189

[11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[12] Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*. 1137–1143.

[13] Alice Newton, Ana C. Brito, John D. Icely, Valérie Derolez, Inês Clara, Stewart Angus, Gerald Schernewski, Miguel Inácio, Ana I. Lillebø, Ana I. Sousa, Béchir Béjaoui, Cosimo Solidoro, Marko Tosic, Miguel Cañedo-Argüelles, Masumi Yamamuro, Sofia Reizopoulou, Hsiao-Chun Tseng, Donata Canu, Leonilde Roselli, Mohamed Maanan, Sónia Cristina, Ana Carolina Ruiz-Fernández, Ricardo F. de Lima, Björn Kjerfve, Nadia Rubio-Cisneros, Angel Pérez-Ruzafa, Concepción Marcos, Roberto Pastres, Fabio Pranovi, Maria Snoussi, Jane Turpie, Yurii Tuchkovenko, Brenda Dyack, Justin Brookes, Ramunas Povilanskas, and Valeriy Khokhlov. 2018. Assessing, quantifying and valuing the ecosystem services of coastal lagoons. *Journal for Nature Conservation* 44 (2018), 50–65. https://doi.org/10.1016/j.jnc.2018.02.009

[14] Derek A Pisner and David M Schnyer. 2020. Support vector machine. In *Machine learning*. Elsevier, 101–121.

[15] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. 2019. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9, 3 (2019), e1301.

[16] Hassan Ramchoun, Youssef Ghanou, Mohamed Ettaouil, and Mohammed Amine Janati Idrissi. 2016. Multilayer perceptron: Architecture optimization and training. (2016).

[17] Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke K Nandi. 2018. Credit card fraud detection using AdaBoost and majority voting. *IEEE access* 6 (2018), 14277–14284.

[18] Jonas Ranstam and JA Cook. 2018. LASSO regression. *Journal of British Surgery* 105, 10 (2018), 1348–1348.

[19] A. Razinkovas, Z. Gasiūnaitė, P. Viaroli, and J. M. Zaldívar. 2008. Preface: European lagoons—need for further comparison across spatial and temporal scales. *Hydrobiologia* 611, 1 (01 Oct 2008), 1–4. https://doi.org/10.1007/s10750-008-9463-4

[20] Javier Senent-Aparicio, Adrián López-Ballesteros, Anders Nielsen, and Dennis Trolle. 2021. A holistic approach for determining the hydrology of the mar menor coastal lagoon by combining hydrological & hydrodynamic models. *Journal of Hydrology* 603 (2021), 127150. https://doi.org/10.1016/j.jhydrol.2021.127150

[21] Aakanksha Sharaff and Harshil Gupta. 2019. Extra-tree classifier with metaheuristics approach for email classification. In *Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018*. Springer, 189–197.

[22] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. 2012. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics* 4, 3 (2012), 275–294.

[23] Zheng Zhang, Zhihui Lai, Yong Xu, Ling Shao, Jian Wu, and Guo-Sen Xie. 2017. Discriminative elastic-net regularized linear regression. *IEEE Transactions on Image Processing* 26, 3 (2017), 1466–1481.