NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

MANUEL BERNARDO RIBEIRO DE OLIVEIRA

BSc in Computer Science

EVALUATING THE ROLE OF ENVIRONMENTAL VARIABLES ON SHELLFISH BIOTOXIN CONTAMINATION VIA SUPERVISED LEARNING

MASTER IN COMPUTER SCIENCE

NOVA University Lisbon
December, 2023

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
COMPUTER SCIENCE

EVALUATING THE ROLE OF ENVIRONMENTAL VARIABLES ON SHELLFISH BIOTOXIN CONTAMINATION VIA SUPERVISED LEARNING

MANUEL BERNARDO RIBEIRO DE OLIVEIRA

BSc in Computer Science

Adviser: Susana Maria Nascimento
Assistant Professor, NOVA University Lisbon

Co-adviser: Marta Belchior Lopes
Assistant Researcher, NOVA University Lisbon

Examination Committee

Chair: Sérgio Marco Duarte
Assistant Professor, NOVA University Lisbon

Rapporteur: Pedro Mariano
Assistant Researcher, ISCTE-IUL

Adviser: Susana Maria Nascimento
Assistant Professor, NOVA University Lisbon

MASTER IN COMPUTER SCIENCE

NOVA University Lisbon
December, 2023

Evaluating the role of environmental variables on shellfish biotoxin contamination via supervised learning

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.10.10) [41].

To my family.

Acknowledgements

I would like to express my deepest gratitude to my advisers, Professor Susana Nascimento and Doctor Marta Lopes, for their ceaseless guidance during the development of this dissertation. Their constant help, availability and desire for excellence were crucial for the work accomplished.

I would like to show my appreciation to project "MATISSE: A Machine Learning-Based Forecasting System for Shellfish Safety" (DSAIPA/DS/0026/2019) for the funding and to the responsible team for all the help and resources they provided.

Additionally, I would like to thank my family and friends for all the support and wisdom given to me during my path that, definitely, contributed to my success.

Lastly, I would like to show my gratefulness to my girlfriend, Carolina, that, through all the support and motivation provided, immensely elevated the produced work.

"You cannot teach a man anything; you can only help him discover it in himself." (Galileo)

Abstract

The production and harvest of shellfish is threatened by harmful algal bloom events that can contaminate these filter-feeding organisms with marine biotoxins. Several studies have been carried out on this topic, but harmful algal blooms are a complex phenomena that still require further investigation to better understand its occurrence and its impact on shellfish contamination. After studies in the Portuguese mainland coast regarding shellfish contamination by marine biotoxins and coastal upwelling recognition through remotely sensed sea surface temperature images, this dissertation aims at broadening the knowledge on this area by studying shellfish contamination in several shellfish production regions in the Portuguese coast and assessing the role on this phenomenon of several environmental drivers including meteorological, hydrodynamic, water properties and coastal upwelling variables. Combining data acquired from previous works and partner institutions, this dissertation focuses on developing an appropriate experimental protocol capable of constructing several machine learning models capable of predicting shellfish contamination, exploring different approaches and algorithms. The work developed included an initial data preprocessing and analysis stage, that merged the data from distinct spatio-temporal sources and selected the best regions and variables. The models for shellfish contamination prediction were developed considering both classification and regression approaches, exploring the predictions as contamination classes or as biotoxin concentration levels. The algorithms used in this work, Random Forest and Support Vector Machine, were selected due to adequacy of their characteristics to the problem and past uses in the literature. The classification approach proved the most successful, correctly predicting most shellfish contamination data cases across the different zones. The inclusion of environmental variables in various combinations proved beneficial for certain models and regions.

Keywords: Shellfish Contamination, Environmental Drivers, Random Forest, Support Vector Machine

Resumo

A produção de bivalves é ameaçada pela proliferação de algas nocivas, que podem contaminar estes animais filtradores com biotoxinas marinhas. Vários estudos têm explorado este tópico, porém, é um fenómeno complexo, cuja ocorrência e impacto na contaminação de bivalves não são completamente compreendidos. Seguindo estudos realizados na costa Portuguesa relativos à contaminação de bivalves e ao reconhecimento do afloramento costeiro através de imagens remotamente obtidas da temperatura da superfície do mar, esta dissertação procura aumentar o conhecimento nesta área através do estudo da contaminação de bivalves em diversas regiões de produção de bivalves na costa portuguesa e avaliar o impacto de diferentes factores ambientais incluindo váriaveis meteorológicas, hidrodinâmicas, de propriedades da água e do afloramento costeiro. Combinando os dados adquiridos de trabalhos anteriores e instituições parceiras, esta dissertação foca-se em desenvolver um protocolo experimental apropriado capaz de construir vários modelos de aprendizagem automática capazes de prever a contaminação de bivalves, explorando diversas abordagens e algoritmos. O trabalho desenvolvido incluiu uma fase inicial de pré-processamento e análise de dados, que juntou dados de diferentes fontes espaciotemporais e selecionou as melhores regiões e variáveis. Os modelos de previsão da contaminação de bivalves foram desenvolvidos tendo em consideração abordagens de classificação e regressão, explorando as previsões como classes de contaminação ou como níveis de concentração de biotoxinas. Os algoritmos utilizados neste trabalho, Floresta Aleatória e Máquina de Vetores de Suporte, foram selecionados devido à adequação das suas características ao problema e ao seu uso na literatura explorada. A abordagem de classificação demonstrou ser melhor sucedida, prevendo correctamente a maioria das amostras de contaminação de bivalves nas diferentes regiões. A introdução de variáveis ambientais em diversas combinações provou ser benéfico para certos modelos e regiões.

Palavras-chave: Contaminação de Bivalves, Factores Ambientais, Floresta Aleatória, Máquina de Vetores de Suporte

Contents

List of Figures

List of Tables

List of Listings

Acronyms

ANN Artificial Neural Network *(pp. 6–11, 23)*
ARIMA Autoregressive Integrated Moving Average *(p. 8)*
ARIMA-DBN Autoregressive Integrated Moving Average - Deep Belief Network *(p. 8)*
ASP Amnesic shellfish poisoning *(pp. 3, 16)*
AST Amnesic shellfish toxins *(pp. 2, 3)*
AUC Area Under The Curve *(pp. 28, 29, 60, 61, 65)*
AZA Azaspiracid Toxins *(pp. 2, 3)*

BOD Biochemical Oxygen Demand *(p. 11)*
BT Boosted Tree *(p. 10)*

CART Classification and Regression Tree *(p. 25)*
chl-a Chlorophyll-a *(pp. 7–11, 14–16, 18)*
COD Chemical Oxygen Demand *(p. 11)*
CUTI Coastal Upwelling Transport Index *(p. 14)*

DA Domoic acid *(p. 16)*
DBN Deep Belief Network *(p. 8)*
DBSCAN Density-based spatial clustering of applications with noise *(p. 42)*
DO Dissolved Oxygen *(pp. 8, 15)*
DSP Diarrheic shellfish poisoning *(pp. 3, 4, 16, 34, 35, 38–40, 43–46, 50, 58–61, 65, 67, 69, 70, 74, 79, 80, 83, 84, 87–90, 92–96, 100–102)*
DST Diarrheic shellfish toxins *(pp. 2, 3)*

ED Euphotic Depth *(p. 9)*

FDA Food and Drug Administration *(p. 11)*
FN False Negative *(pp. 56, 57, 67–69, 72)*
FP False Positive *(pp. 56, 57, 67, 68, 72, 74)*

GA Genetic Algorithm *(p. 7)*
GAMs Generalized Additive Models *(p. 16)*
GP Genetic Programming *(p. 7)*

HAB Harmful Algal Bloom *(pp. 2–4, 6, 8–17)*
HAEDAT Harmful Algal Event Database *(p. 3)*
HTRBF Heavy-Tailed Radial Basis Function *(p. 10)*
HWP Hydrodynamic and Water Properties *(pp. 3, 4, 34, 39, 40, 43, 44, 53, 63, 74, 76, 84, 87, 95, 96, 100–102)*

IPMA Instituto Português do Mar e da Atmosfera *(pp. 3, 34, 38)*
IRF Iterative Random Forest *(pp. 9, 25)*

KPCA Kernel Principal Component Analysis *(p. 10)*

LSTM Long Short-Term Memory *(pp. 8, 11)*

MAE Mean Absolute Error *(pp. 29, 49, 83, 84, 87, 88)*
mCART Modified Classification and Regression Tree *(p. 25)*
MDA Mean Decrease Accuracy *(p. 26)*
MEI Multivariate El-Niño Southern Oscillation Index *(p. 15)*
MI Mutual Information *(p. 10)*
ML Machine-learning *(p. 6)*
MLD Mixed Layer Depth *(p. 16)*
MLP Multilayer Perceptron *(pp. 6–8)*
MRF Meteorological Risk Factor *(p. 13)*
MSE Mean Squared Error *(pp. 29, 49)*

NOAA National Oceanic and Atmospheric Administration *(p. 14)*

OOB Out of Bag *(p. 24)*

PAR Photo-synthetically Active Radiation *(pp. 11, 16)*
PCA Principal Component Analysis *(p. 14)*
PR Precision-Recall *(p. 29)*
PSP Paralytic shellfish poisoning *(pp. 3, 11, 16)*
PST Paralytic shellfish toxins *(pp. 2, 3, 11, 13)*

QRF Quantile Random Forest *(pp. 8, 9, 26)*

R^2 Coefficient of Determination *(pp. 29, 49, 53, 88)*

RBF Radial Basis Function *(pp. 10, 27)*

RF Random Forest *(pp. 3, 8–11, 15, 16, 20, 21, 23–26, 31, 32, 49–53, 56, 59, 60, 65, 68, 73, 74, 78–80, 84, 89, 92–94, 96, 100, 101)*

RFE Recursive Feature Elimination *(p. 20)*

RMSE Root Mean Squared Error *(pp. 29, 49, 83, 84, 87, 88)*

RNN Recurrent Neural Network *(p. 8)*

ROC Receiver Operating Characteristics *(pp. 28, 29, 50)*

SE Southeast *(p. 11)*

SSH Sea Surface Height *(p. 15)*

SST Sea Surface Temperature *(pp. 2, 8, 9, 11, 13–18, 41, 42)*

SVM Support Vector Machine *(pp. 3, 9, 10, 20, 23, 26–28, 31, 32, 39, 49, 51, 54, 56, 68, 73, 74, 78, 79, 95, 96, 100–102)*

SVR Support Vector Regression *(pp. 28, 31, 32, 39, 49, 54, 80, 87, 89, 92, 93, 95, 96, 99, 101, 102)*

SW Southwest *(p. 9)*

TN True Negative *(pp. 56, 57, 68, 72)*

TNR True Negative Rate *(pp. 29, 56–58, 61, 63, 66–69, 72–74, 95, 96)*

TP True Positive *(pp. 56, 57, 67–69, 72–74)*

TPR True Positive Rate *(pp. 28, 29, 56–58, 61, 63, 67–69, 72–74, 95, 96)*

UI Upwelling Index *(pp. 9, 14, 15, 17)*

Uo Zonal Component of Water Transport *(p. 15)*

YTX Yessotoxins *(pp. 2, 3, 16)*

1 Introduction

1.1 Motivation

The shellfish production and harvesting sector is affected by water contamination events, which are not fully understood and are associated with complex relations with other phenomena, namely, coastal upwelling. Nowadays, large amounts of data regarding these occurrences and factors that can contribute to them are available from several monitoring programs and sources. When combined with robust mathematical and computational models, capable of processing this quantity of data it is possible to better understand the intricate dynamics around shellfish contamination.

This dissertation aims to assist food safety authorities manage suspensions on shellfish harvesting and production by providing a deeper understanding on how upwelling events are connected to shellfish contamination. This information can be added to existing monitoring programs, expanding the information available to the authorities, strengthening their decision making ability, increasing effectiveness, minimizing risk and optimizing suspension times.

This work proves itself useful to shellfish producers as it can help manage production and minimize financial losses, by improving data on shellfish contamination in their regions. Additional information can empower these producers by helping them manage their production and shellfish stock and foresee future bans and the lift of current ones. This work is inserted in project Matisse: A Machine Learning-Based Forecasting System for Shellfish Safety (DSAIPA/DS/0026/2019), a project funded by the Portuguese Foundation for Science and Technology, intended to support decision making in Public Administration and the shellfish production sector.

1.2 Problem Description

Shellfish species are filter-feeding organisms and, therefore, its production is affected by the surrounding water quality. This can be affected by algae proliferation, with the aggravating factor that some of these species can produce toxins that lead to a contaminated

environment and, subsequently, cause contamination in shellfish. An increase in algae abundance often stems from Harmful Algal Bloom (HAB) [26], a phenomenon in which algae populations rapidly increase and create a negative impact on the surrounding wildlife by depriving the environment of oxygen.

In the west coast of the Iberian peninsula HAB occurrences are frequently associated with coastal upwelling [48]. This event occurs when cold waters from the deep ocean are brought closer to shore, pushing coastal waters further into the ocean, altering the vertical profile of the water column. This occurs seasonally every year in the continental Portuguese coast, from late February to early September. The phenomenon results in a stronger nutrient presence near the shore creating an ideal environment for a prosperous wildlife development. Most maritime economic activities thrive in coastal upwelling zones, namely shellfish harvesting. These nutrient rich waters are, however, a common location for the appearance of HAB which threatens the local and surrounding wildlife. Since coastal upwelling involves a high intensity of maritime currents, these newly formed HAB can travel large distances and endanger other environments [61]. Recent work has been developed on the automatic recognition of coastal upwelling from Sea Surface Temperature (SST) sattelite images via novel spatial clustering [53, 52] as well as unsupervised spatio-temporal clustering framework for the spatio-temporal analysis of coastal upwelling [44, 55].

The affected shellfish represent a health risk to consumers and, consequently require close monitoring from food safety agencies in order to avoid a public health crisis. These organizations are tasked with ensuring toxin concentrations are within thresholds established by European law, present in Table 1.1 for toxic algae concentration, by performing frequent sample inspections. The detection of toxin levels surpassing the designated limits implies a harvesting cessation in the affected area until values are deemed appropriate. The pause in production causes uncertainty among producers and great economic losses for the sector.

Table 1.1: Toxin producing algae concentration limits in the water. According to [1]

Toxin group	**Algae Species**	**Concentration limit** *cells/Liter*
Paralytic shellfish toxins (PST)	*Gymnodinium catenasum*	>1.500
	Alexandrium spp.	>1.500
	Marine Cianobacterias	20.000.000
Amnesic shellfish toxins (AST)	*Pseudo-nitzchia spp.*	200.000
	Pseudo-nitzchia	1.000.000
Diarrheic shellfish toxins (DST)	*Dinophysis spp.*	500
	Prorocentrum spp.	1.000
Azaspiracid Toxins (AZA)	*Azadinium spp.*	1.000.000
Yessotoxins (YTX)	*Gonyaulax spinifera*	1.000.000

Ingesting contaminated shellfish can cause a range of poisoning syndromes, such

as Paralytic shellfish poisoning (PSP), Amnesic shellfish poisoning (ASP) and Diarrheic shellfish poisoning (DSP). Generally in Portuguese waters, *Gymnodinium catenasum* are associated with producing PST, *Pseudo-nitzchia spp.* with producing AST, *Dinophysis spp.* with producing DST, *Azadinium spp.* with producing AZA and *Gonyaulax spinifera* with producing YTX [20]. According to the Harmful Algal Event Database (HAEDAT) [32] there were 689 HAB related events from 1985 until 2022 reported in Portugal, with the most prevalent one being of the DSP type. There were also PSP and ASP type HAB reported events but no AZA related event has been recorded. The HAB that produced toxins from type DSP were formed mostly during the upwelling season, while HAB associated with producing toxins from type PSP and ASP were more common during the relaxation of upwelling.

Mainland Portugal is divided into 39 shellfish production areas, 13 of which are coastal areas and the remaining 26 are lagoons or estuaries. For each production area, water and shellfish samples are collected and monitored weekly by Instituto Português do Mar e da Atmosfera (IPMA). The collected shellfish samples include an indicator species, that accumulate biotoxins at a higher rate. The water samples are tested for cell concentration per liter of algae species associated with producing marine biotoxins. In addition to this test, samples of existing species are collected and tested for marine biotoxin concentrations. If the tested indicator species (i.e., the mussel *Mytilus Galloprovincialis*) surpasses the regulatory limit of any toxin, the harvest is suspended in the region. As different species accumulate toxins at varying rates, when the zone is closed, other species can be tested, and if their toxins are within the legal limits its harvest is allowed. In coastal areas, upwelling effects are more prominent and its impact on shellfish contamination can be better studied [76], when compared to other regions, such as estuaries, that can be impacted by other external phenomena, namely river discharges [74]. Due to these characteristics, this dissertation will focus on the coastal areas.

1.3 Objectives

The main goal of this dissertation is to develop an architecture based on an experimental protocol with two well-known algorithms, Random Forest (RF) and Support Vector Machine (SVM), to evaluate the role of environmental variables on shellfish contamination. To build these models, data used in a previous work on shellfish contamination forecasting [17, 16] will be used in conjunction with data from a work on automatic recognition of coastal upwelling [44, 55] and new Hydrodynamic and Water Properties (HWP) data. These models will be developed, in parallel, as classification and regression tasks, each with distinct datasets, to determine the best approach for this work. The classification datasets, will include a binary outcome variable classifying shellfish samples as contaminated or not contaminated based on the DSP concentration legal limit referred in Table 1.2. On the other hand, for datasets used in regression, an outcome variable regarding the predicted DSP concentration value of the shellfish samples will be used.

Table 1.2: Toxin levels in shellfish according to European Union Regulations n.º 853/2004, n.º 786/2013, n.º 2021/1374 and n.º 2021/1709.

Poisoning syndrome	Associated toxins	Legal limits
PSP	STX and analogs	800 µg kg^{-1}
ASP	DA	20 mg kg^{-1}
DSP	OA, Dinophysis and PTX	160 µg kg^{-1}
AZP	AZA	160 µg kg^{-1}
YTX	YTX	3,75 mg kg^{-1}

From the dataset obtained from the previous works, the most important features will be identified through an extensive review of relevant studies. After this step, the data will be split into train, validation and test datasets that will be used to build the models. Finally, their performance will be evaluated with the use of selected evaluation metrics.

By considering environmental data from different spectres, this dissertation will research the benefits of exploring a broad range of variables when studying shellfish contamination. The role and importance of this data will be evaluated through the relevance of its features to the models as well as their contribution to the results.

1.4 Contributions

The work developed in this dissertation proved capable of predicting DSP concentration allowing the study of shellfish contamination by marine biotoxins. The two approaches, regression and classification, followed for the models' construction, provided different perspectives regarding the studied problem, allowing better evaluation of the results. Considering the algorithms tested in each approach, similar results were obtained by the best models in every selected region. This similitude between results was also present when comparing the best regression predictions converted into binary contamination classes and the classification models' best results. Although the best results were mostly obtained using only DSP concentration data with the most recent DSP value almost always proving to be the most important variable in the predictions, HWP and upwelling variables presented promising results prompting further research to evaluate their full potential.

1.5 Organization of the Document

Following this introductory chapter, the subsequent chapters are organized with the following structure. In Chapter 2, related works focused on topics such as prediction of HAB and shellfish contamination and the role of environmental variables on HAB are discussed in order to gain necessary knowledge on the key phenomena studied in this dissertation. This chapter finalizes with an analysis on the most prominent features used in the reviewed studies. The fundamental theoretical concepts required to understand

the developed practical work are presented in Chapter 3, where key notions as the chosen algorithms and common evaluation metrics are detailed. In Chapter 4, the proposed methodology is presented, followed by a detailed description of the datasets' construction and analysis. This chapter ends with the models' construction steps delineated and their hyperparameter tuning phase analysed. The results obtained by these models are presented and compared in Chapter 5. The analysis starts by comparing models of each algorithm and approach using different combinations of variables, followed by a comparison of algorithms within each approach and ending with an overview of the best results of each approach. Finally, Chapter 6 contains the conclusions obtained from all the developed work and its achieved results while also proposing future steps to continue to improve the research pursued in this dissertation.

2

HAB and Shellfish Contamination: Related Work

In this chapter, firstly, a review on research done on HAB, shellfish contamination and the role of environmental variables on these events will be presented. In the end, an analysis on the most commonly used features in the reviewed literature is performed.

2.1 HAB and Shellfish Contamination Forecasting

HAB occurrence and the associated shellfish contamination have been an important scientific topic throughout the years due to the negative impact they have on coastal areas and other bodies of water.

2.1.1 Harmful Algal Blooms

One of the main fields of research on this topic has been forecasting HAB and variables associated with it along with studying their dynamics. With its complex nature still not fully understood, HAB forecasting can be treated as a nonlinear regressive problem. Machine-learning (ML) models have gained popularity across the years when compared to traditional models, such as process-based models and statistically-based models, with many studies using Artificial Neural Network (ANN) due to it's capability of dealing with missing values among data and detecting non-linear, complex and dynamic relations across a range of parameters. These models require a large set of data in order to obtain a good performance and are hard to extract information from.

In an early study, Maier et al. (1998) [42] used Multilayer Perceptron (MLP) models with one hidden layer to forecast *Anabaena spp.* concentrations four week in advance in the Murray-Darling river system. It focused on evaluating the impact of eight selected variables on cyanobacteria abundance and the benefits gained from using lag-time in these variables. To evaluate the parameters' impact, a stepwise modelling approach was used. Starting with single variable models, and incrementally adding the best performing ones until new additions didn't improve the performance. Temperature was the best predictor

in the single variable models, and the introduction of lag-time to the variables proved to have a meaningful advantage. The best model overall was obtained using temperature, flow and colour, being able to forecast the start and duration of intense *Anabaena spp.* abundance, presenting, however, lower accuracy in extreme abundance cases.

In another study, Recknagel et al. (2002) [65], after conducting a case study comparing past ANN based approaches [64], compared the performance of ANN and Genetic Algorithm (GA) models to predict blue-algal abundance in lake Kasamiura, Japan seven days ahead. The models using a range of biological, physical and chemical predictors managed to predict the overall trend of the bloom but failed for extreme peaks.

Another ANN based approach was developed by Lee et al. (2003) [36] proposing a MLP model to forecast algal blooms in Hong Kong. Despite having a wide range of data available, containing several meteorological and physical-chemical features, the best models were always found by simply using previous Chlorophyll-a (chl-a) or cell concentrations.

In a study conducted in Andalucía, by Velo-Suárez and Gutiérrez-Estrada (2007) [75] in order to better understand *Dinophysis acuminata* bloom dynamics that were affecting the local shellfish production, proposed several MLP models using *Dinophysis acuminata* abundance values. These models were developed for seven days ahead forecasting, using data from different regions in the area, to evaluate spatial variations between regions. As previously reported in other studies, the models were able to identify and predict the overall variations of abundance, losing accuracy when dealing with extremes values, specially with the lower end.

In a more recent study in Alfacs bay, Guallar et al. (2016) [24] developed several short-time forecast MLP models used for detection through presence classification and for forecasting abundance, focusing on two common species in the region, namely the microalgae group *Pseudo-Nitzchia* and the dinoflagellate group *Karlodinium*. The first model was used to evaluate the presence of the species, filtering then the samples with a positive presence to use on the second model. Both models were developed to forecast values one week in advance and predictors were selected from a robust set of biological and environmental variables. Both models achieved a good performance and the best results were obtained when using both types of variables. Presence-detection models achieved a better accuracy than abundance forecasting models, this was justified by having more data available and a higher complexity that allowed for better model performance. A decrease in accuracy occurred for *Karlodinium* presence models when using the most recent data, indicating that some environmental changes may be happening in the Catalan ecosystem. Differences among the size of data required to achieve a good performance were identified between species, with *Pseudo-Nitzschia* needing more than fifteen years of data, while *Karlodinium* absence-presence model was able to obtain a high accuracy using only five years of data.

In another study, Muttil and Chau (2006) [49] developed a MLP with one hidden layer model and a Genetic Programming (GP) model for real-time one-week ahead algal

bloom prediction in Tolo Harbour, Hong Kong. The forecasting was done through chl-a abundance as it was considered an indicator of algal biomass. For the MLP model, a lag-time of one week was used and an analysis of the network weights was performed to identify the most influential predictors, with chl-a previous concentration being the most impactful. Other tests carried out with biweekly data to test for interpolation effect supported the same results. Overall the model using chl-a and weekly data achieved the best performance.

Similarly, Lee and Lee (2018) [37] compared MLP, Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) [22] models using three hidden layers for forecasting HAB one week ahead in the four biggest rivers in South Korea. For the comparison all models used nine selected predictors, which were ranked by importance using backward elimination, with temperature and Dissolved Oxygen (DO) ranked as the most important. The LSTM model performed better than the other two models, with MLP being better than the LSTM in some specific cases but failing to track the overall pattern in other cases, struggling with extreme values.

To determine algal bloom risk daily in Tolo Harbour, Guo et al. (2020) [25] proposed a MLP model with one hidden layer. Based on the Stability Theory [80, 79], that referred that for algal bloom occurrence a stable water column with an euphotic layer on top is needed. The model was used to forecast SST and vertical temperature and salinity differentials for the next day. The predicted values together with the latest nutrient data, were used to assess the daily algal bloom risk. A second model was proposed using only physical predictors for when the daily nutrient data was not available, having a worse performance than the previous model but proving useful as a backup for periods with missing data.

Autoregressive models have also been used for HAB forecasting, with Qin et al. (2017) [63] proposing a hybrid Autoregressive Integrated Moving Average - Deep Belief Network (ARIMA-DBN) model to predict HAB occurrence two months ahead in the Zhejiang province, China. The model consisted of a first Autoregressive Integrated Moving Average (ARIMA) [8] layer used to forecast the environmental predictors, while Deep Belief Network (DBN) [31] served to capture the complex nonlinear relationship between them and the HAB abundance, as the ARIMA is not capable of detecting nonlinear patterns. The model achieved great accuracy and when compared to a simple DBN and to a different ANN, outperformed the other two.

In more recent years, RF [9] models have also been proposed for HAB forecast. RF models have several advantages including not requiring initial assumptions, having an easier interpretation of outputs and balancing imbalanced data.

Asnaghi et al. (2017) [2] proposed a Quantile Random Forest (QRF) [47] regression model to forecast *Ostreopsis cf. ovata* abundance in the Ligurian coast, using weekly meteorological variables. In a previous study Asnaghi et al. (2012) [3] observed that seawater temperature hydrodynamics had a very high impact on bloom dynamics in the region and were of extreme importance to better understand the phenomenon. The most influential variables of the model were determined to be SST and day of the year,

strengthening the previous study conclusions. The biggest benefits from the QRF model was featuring a large flexibility due to the fact that predictions can be generated at chosen quantiles, allowing for choosing the risk level on the approach, being overall able to accurately predict abundance, with some overestimation for lower values.

In a recent study, Derot et al. (2020) [18] proposed a RF method to classify HAB in Lake Geneva, which had been suffering from eutrophication in the past decades, making it more prone to proliferation of HAB. Four classes were created for classification using K-Means [29] to categorize cyanobacteria intensity, the model used these classes to classify concentrations of *Plankothrix rubescens*. The overall results weren't great but it was able to classify the intensity of harmful *cyanobacteria* abundance in Lake Geneva over a year scale.

Cheng et al. (2021) [13] proposed two Iterative Random Forest (IRF) [5] models, one was used to identify key factors and interactions among abiotic variables and HAB occurrences, with the other being used to identify physical, chemical and biological drivers on microbial abundances. The results highlighted for the first case, phytoplankton abundance in response to coastal conditions and inland nutrient fluxes, and for the second case, it identified abundances of some algae species and chl-a to be dominant drivers on the microbial concentrations.

Several SVM [7] models have also been developed in more recent years. Compared to ANN, SVM shows certain advantages as performing well with smaller volumes of data and assuring a global optimal solution, with its biggest disadvantage being its running time, taken severely longer than other approaches.

In a recent study in Charlotte County, southwest Florida, Izadi et al. (2021) [33] compared XGBoost, RF and SVM models on forecasting the start of HAB. This study focused on understanding the impact and optimal lag times of variables related to the start of blooms,focusing on *Karenia brevis*. An increase in performance was obtained in all three models when using several days of lagged input, being limited by available data to three days. All models achieved a good performance, with XGBoost achieving slightly better performance, and having the best results for forecasting in eight days in advance. SVM showed best performance for seven days ahead forecast, with RF also performing the best for eight days ahead. chl-a, SST, Secchi disc depth, and Euphotic Depth (ED) were found to be the most impactful predictors in this model, with ED, SST, and chl-a being the most significant variables among all models.

Vilas et al. (2014) [76] proposed a SVM model for predicting HAB of *Pseudo-Nitzchia* in the Southwest (SW) Galician coast. Focusing in the four main coastal inlets of the region and using eight years of environmental data, from zones where upwelling impact is greater and the coastal inlet natural effects are lower. The data was separated in four classes according to *Pseudo-Nitzchia* cell concentration and an Upwelling Index (UI) was calculated using wind data by the Bakun's method [4] for the sampling day and also for the previous days considering the phenomenon evolution. Analysing the results the authors didn't find any direct relations between the studied parameters and the abundance of *Pseudo-Nitzchia* blooms, however patterns related to their spatial and temporal distribution

were found. It was observed that blooms frequency and duration varied across the four coastal inlets and this could be due to their topography and wind conditions. Looking at the data, it was also possible to identify seasonality among bloom detection, with most of them occurring during upwelling season. Overall the SVM achieved a great accuracy, specially for the bloom/no bloom models.

Ribeiro and Torgo (2008) [66] did a comparative study on predicting algae blooms in Douro River, using SVM, ANN and regression trees models, trained with algae species concentrations and some physical-chemical parameters. Due to the presence of algal blooms being an uncommon event associated with extreme algae concentration that contrast with average abundance values, the authors [72] developed a error statistic to evaluate the accuracy of the developed models in predicting these scarce events. Evaluating the models performance using the developed error statistic showed that no model achieved particular good results, but SVM provided a better accuracy in predicting extreme abundances, which are associated with blooms, when compared with the other two models.

Gokaraju et al. (2011) [23] proposed a SVM classification model to better understand the spatial-temporal dynamics of HAB blooms occurrence. The proposed model was developed using a Kernel Principal Component Analysis (KPCA) combined with a SVM using the Heavy-Tailed Radial Basis Function (HTRBF) kernel. Using KPCA and wavelet for feature extraction, Mutual Information (MI) criterion for measuring feature importance and lagged inputs, the SVM model was shown robust and capable of classify between HAB and non-HAB blooms, with chl-a related features showing as the most influential.

To best predict chl-a concentrations in Tolo Harbour, Hong Kong, Li et al. (2014) [38] compared a SVM model using a Radial Basis Function (RBF) kernel to several ANN models. All the models were built using nine lagged inputs and tested with seven and fourteen days lead windows. It was observed that either for the seven-day lead model and the fourteen-day lead model, SVM outperformed the remaining four ANN related models, with these showing worse generalization power and considerably worse overall results. The only downside of SVM compared to the other proposal was the runtime where it took several times the amount of the others.

In another study in a eutrophic environment, Harris and Graham (2017) [28] analysed the temporal cycles of cyanobacteria blooms in Cheney Reservoir, Kansas. Four types of models were developed and compared, namely, SVM, RF, Boosted Tree (BT) and Cubist modelling, to predict cyanobacteria, microcystin and geosmin abundance. Due to a high number of variables, the ones with a high correlation coefficient were removed, as were the ones with a considerable amount of missing values. It was observed that generally, predicted values were greater in late Summer and early Autumn, microcystin abundance seemed to precede peaks of cyanobacteria. For cyanobacteria abundance, SVM,RF and BT showed the best results, with SVM obtaining slightly better accuracy than the other two. These models identified elevation and chl-a as the most important features, with temperature, phosphorus, iron and current season as also relevant features. Nevertheless,

it was noted that all the models performed poorly when dealing with the higher spectre of values, being justified as a possible reason, the use of seasonal variables for a phenomenon that didn't show any seasonality in the study region.

Yajima and Derot (2018) [81] proposed a RF model to forecast chl-a concentrations in a reservoir and a lake in Japan. In these water bodies eutrophication is an increasing worry and has been linked to HAB occurrences. The two water bodies showed significant differences between them, as the Reservoir was man made, had shorter data, while the lake had larger data, and a higher complexity resulting from seawater mix and river discharges. To evaluate the performance of the model, a sliding window was used, adjusting the lead-time and lag-time. The model didn't perform well in forecasting chl-a concentrations in either water body but was still deemed useful for other tasks as quality control. When comparing between the two different water bodies, the number of features and the size of data was deemed more important than the overall differences between the bodies, and Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) were considered to be the most important predictors.

2.1.2 Shellfish Contamination

Besides the relevance of predicting HAB, an emerging research topic is directly forecasting shellfish contamination by predicting toxins concentration, as HAB events not always lead to shellfish harvesting interdictions and the accumulation-elimination dynamics varies across species. This field has been seldomly explored and more research can expand knowledge in this area.

In one of the available studies, Harley et al. (2020) [27] proposed a RF classification model to determine environmental drivers and provide a short-term forecast of PSP contamination in Southeast (SE) Alaska. This region had a lack of previous studies and was characterized by a complex oceanographic structure, being influenced by glacial discharges and coastal upwelling. Available PST concentration, environmental data, such as SST and Photo-synthetically Active Radiation (PAR), meteorological data, such as air temperature, precipitation and wind speed, and an upwelling index were used in the model, to classify toxin concentrations as above or below the Food and Drug Administration (FDA) limit. Evaluating the results, SST, air temperature, current season and salinity ranked as the most important features. Additionally, when a seven-day lag was introduced, it improved the model's performance due to the average ten day difference between the predicted day for the toxin concentration increase and the actual registered date.

In a recent work, Cruz (2022) [17] proposed ANN and autoregressive models to forecast shellfish contamination. The LSTM models performed the best out of all models, achieving great accuracy for one week ahead predictions, with ANN models performing generally better than the autoregressive models. Contrary to what was expected, an increase in variables used did not correlate to a better performance.

2.2 The Role of Environmental Variables on HAB

In several regions across the globe, HAB occurrence is frequently associated with several environmental drivers including meteorological and hydrodynamic variables and phenomenona as coastal upwelling and climate change.

An analysis on red tides including their causes, impacts and control methods was performed by Zohdi and Abbaspour (2017) [83]. Red tides are characterized as algae blooms in marine environments, such as estuaries and coastal regions, that present a coloration, most commonly red or green, observable from the water surface. This phenomenon is associated with a rapid increase in the concentration of algae species, most commonly, diatoms and dinoflagellates. The factors identified as the main contributors to the occurrence of red tides include wind, ocean currents, temperature and nutrient concentration. The main causes of red tides identified in this study are categorized as resulting from human activities or natural occurrences. The former includes activities that lead to the increase in nutrient concentration in the water such as agricultural, industrial, aquaculture and the use of ships. The latter includes the increase of environmental variables namely, water temperature, wind intensity, salinity, rainfall and vitamins concentration or the impact of more complex phenomena such as the El Niño, current systems, coastal upwelling, dust storms and ocean eddies. The impacts of red tides identified in this study include direct and indirect poisoning of aquatics species, humans and other animals and the decrease in water quality. The main methods referred to control red tides include the reduction of nutrient influx caused by the previously referred human activities, the introduction of phytoplankton feeding organisms and the chemical treatment of the water.

The climate change impact on HAB was analysed and studied by Wells et al. (2015) [77] through an extensive review of the available literature relating the two phenomena. Climate changes lead to variations in temperature, irradiance and wind, which are key variables to HAB occurrence and development, impacting HAB dynamics that can lead to variations in HAB trends. This study focused on analysing the possible direct effects of climate change related variables to HAB. Temperature is regarded as a key factor for phytoplankton dynamics with its increase, resultant from climate change, leading to variations in the available temporal space for phytoplankton development depending on region and climate. Additionally, the increase in water temperature correlates to the increase in growth rates of the phytoplankton, affecting differently the stages of its development. The biogeographical distribution of the HAB species can be further expanded with the increase in water temperature in higher latitude regions. Climate change is expected to increase water stratification, causing a change in nutrient accessibility. Although this variation does not lead to an increase in HAB, it was possible to identify that certain HAB species thrive under an increased stratification and that this change seems to affect HAB dynamics. Some climate change studies predict future variations in cloud prevalence around the globe. These changes impact phytoplankton species' light exposure, possibly leading to changes in their behaviour. This linkage, however, is

seldomly researched, not allowing a clear understanding of the possible impact of this change. Climate change is also associated with an increase in atmospheric CO_2, that through its dissolution on the water leads to ocean acidification. This phenomenon was researched in early studies that predicted a decrease in the water pH and a higher impact on upwelling regions. Nonetheless, it was not possible to identify how this occurrence would affect HAB, with preliminary studies presenting ambiguous results when considering different species. Regarding increases in nutrient influx resulting from climate change, it was not possible to understand how it affects HAB due to the fact that HAB and non-HAB species presented similar dynamics and that the nutrient influx is also resultant from human activity. Most reviewed works in this study were conducted analysing each factor on its own, however, HAB is a very complex phenomenon that, most likely, is influenced simultaneously by multiple factors. Finally, most available studies focused on selected species or regions using different experimental procedures and techniques, that did not allow global conclusions on the impact of climate change on HAB to be drawn.

The role of environmental and oceanographic variables on recently detected *Alexandrium catenella* dinoflagellate blooms was investigated by Condie et al. (2019) [14]. This study focused on the eastern coast of Tasmania, Australia, for the time period between 2012 and 2018, during which, an unprecedented occurrence of blooms was recorded. The data used included PST concentration, SST, water level, current velocity, modelled shelf circulation, rainfall, air temperature, wind speed, river flow data and water temperature and salinity depth profiles. The followed approach analysed the presence of blooms, through PST concentration, and the water column stratification through rainfall, air temperature and wind speed. A new variable denominated Meteorological Risk Factor (MRF) was created to study stratification using the 3 previously referred variables, being calculated monthly through a ratio between a weighted combination of rainfall and minimum air temperature and wind speed. Analysing the considered variables and the blooms' occurrence, the rainfall, air temperature and wind speed were considered as being heavily related to the bloom occurrence, with two common scenarios being frequently present during high PST concentrations. These scenarios included either high rainfall and low wind speed or low minimum temperature and low wind speed. These patterns highlighted a strong correlation between bloom occurrence and the new MRF variable. As previously referred, the three variables composing MRF were associated with stratification, with this being considered the main factor for the formation of coastal blooms. Finally, the MRF was considered promising for future development of a HAB forecasting mechanism.

The importance of water column stratification for dinoflagellate blooms was also studied by Smayda (2002) [70]. In this study, an alternative hypothesis to the relation between these two phenomena was proposed. The classical paradigm implies that vertical stratification of the water column is an essential requirement for dinoflagellate blooms. From the observed results regarding the behaviour of several dinoflagellate species, it was identified that dinoflagellates are able to endure environments with significant turbulence, meaning that stratification can be a parallel phenomenon and not a requirement. The

proposed hypothesis refers that certain dinoflagellate species can develop in frontal areas and later propagate into neighoubouring regions, possibly leading to blooms.

In a recent study, a HAB forecasting model was proposed by Wen et al. (2022) [78]. This model was tested using several environmental variables collected by research ships and maritime buoys and Alexandrium cells concentration collected by National Oceanic and Atmospheric Administration (NOAA) [69] during a 3 years period on the Maine coast in North America. The first step of this study, and the only one considered relevant for this analysis, consisted of selecting the major variables regarding Alexandrium behaviour, removing unimportant variables and reducing the complexity, through a developed Principal Component Analysis (PCA) variation. This method was used to calculate the correlation of the environmental variables to the Alenxandrium data. From the obtained results, phosphate, nitrogen, fluorescence, pressure, chl-a, latitude, depth and silicon were considered the most influential variables.

Coastal upwelling is frequently associated with HAB occurrence, as upwelling cycles are generally characterized by higher nutrient concentrations, where HAB can thrive, with several studies focusing on topics as the relationship between the two or overall understanding of the coastal upwelling phenomenon [35]. Across literature, it is common to construct a UI using wind data or atmospheric pressure. [76, 19, 58], with the two most common being the Coastal Upwelling Transport Index (CUTI) [34] and the Bakun's method.

A study to explore relationships between upwelling spatio-temporal dynamics and HAB events in Oregon was conducted by Tweedle et al. (2010) [74]. The Oregon region is part of the California Current Upwelling System and is impacted by strong winds, switching from winds to the north (downwelling favorable) during the winter and winds to the south (upwelling favorable) during summer. In previous studies an increase of chl-a concentration during upwelling season was already reported [30, 71]. This study used meteorological parameters, extracted from satellites, associated with upwelling, as SST and chl-a and compared multi-annual seasonal cycles of upwelling and chl-a to HAB events and toxins abundance. The Oregon coast was divided in five regions and SST satellite images were compared at different points in each of them, to detect upwelling through variations in measurements done at coastal and oceanic points, with the latitudinal gradients being then used as a measure of upwelling impact. Oceanic SST was observed to follow a stable cycle, reaching minimum values in March, and maximum values during the summer, being overall colder in the north. The start of bloom conditions were observed to start in early spring, with wind data confirming strong upwelling favorable winds in early spring. The satellite based analysis proved to be capable of identifying HAB-prone scenarios, therefore being a useful approach to study the relations between upwelling and HAB.

In a study in the Portuguese coast regarding bloom development in a coastal upwelling affected area, Ferreira et al. (2021) [21] studied the phytoplankton bloom phenology in the western Iberian coast, a region inserted in the Canary Current Upwelling System. This

study used twenty-two years of chl-a satellite data with daily high resolution. The main focus of this work was to investigate spatial and temporal variations on bloom phenology and its main drivers. For this study the area was divided into phenoregions and a RF model was constructed, with feature importance being measured by Drop-column importance. Analysing the data, upwelling effect was evident, with a decrease in oceanic chl-a from spring to summer and an increase in coastal upwelling centers, and blooms tended appear later during the last twenty years. After assessing the main environmental drivers in each region, the west coast region was found to be more complex without any clear main predictor, considering salinity, Multivariate El-Niño Southern Oscillation Index (MEI), Sea Surface Height (SSH) and Zonal Component of Water Transport (Uo) as the most impactful. In the south coast, the predominance of a main driver was also not clear, with salinity, silicon and Uo performing the best. These results in coastal areas contrast with the oceanic regions evidencing the impact of upwelling, identified by differences in bloom occurrence, with blooms of higher frequency and lower duration, matching the upwelling season and the upwelling brief cycles influence.

Similiarly, Palma et al. (2010) [58] studied on the role of upwelling on *Pseudo-Nitzchia* blooms in Lisbon Bay, constructing an UI based on wind data, and evaluating the index values to determine seasonal trends. The index was used together with SST to evaluate bloom development, showing patterns among ocean seasons. For further study, the data was divided by oceanographic periods and mathematical models were constructed to evaluate the relationships between SST and UI with *Pseudo-Nitzchia* abundance at different lag time values for the global and seasonal data. These models found a significant positive relation between SST and previous concentration, and a negative relation to UI, with the global model showing an input lag of five days and the model by season showing a four day lag for Spring and five day lag for Summer. This showed that upwelling intensity can be prejudicial for HAB since high turbulence generates higher water column mix, so an UI influence limit was introduced in the model by season, improving its performance when compared to the general model. Among variables, SST was found to reflect better the annual variations than UI, with phytoplankton cycle having a clear resemblance with the upwelling cycle.

In a study in the south of Portugal, Cardeira et al. (2013) [11] studied the variances in biological and chemical parameters associated with upwelling events in the area. Upwelling is weaker in the south coast when compared to other regions in the Canary system, being more intense around Cape São Vicente and extending some times to Cape Santa Maria and even further. The research was conducted using a research cruise, analysing vertical and horizontal profiles in different regions in the coast. The horizontal fields showed that the chl-a was negatively related with SST and salinity and positively with DO, being at its highest concentrations in upwelling waters. Analysing the vertical profiles at different depths, the upwelling was evident, it was observed that chl-a decreased rapidly with depth and the other parameters showed a similar behaviour to the horizontal profiles. These results indicate that upwelling has a strong impact on nutrient intake and

is stronger in Cape São Vicente.

In the same region, Lima et al. (2022) [40] recently studied the spatial-temporal variability patterns and phenology of potentially toxigenic phytoplankton species in coastal waters of southern Portugal during a six year period. Focusing on two groups of algae associated with important HAB events in the region, namely diatoms and dinoflagellates and in its associated drivers, Generalized Additive Models (GAMs) models were developed to identify key predictors for HAB species associated with DSP and ASP syndromes in each area. These models proved themselves unuseful, not being able to provide much information, nevertheless, PAR, chl-a, Mixed Layer Depth (MLD) and SST were identified as the most influential, although upwelling intensity was not selected for *Pseudo-Nitzchia* which was more frequent on coastal areas that had a higher intensity of upwelling, more present during spring and summer matching the upwelling season and had more intense and long blooms, contrasting with *Gymnodinium catenatum* blooms that were generally delayed when compared with *Pseudo-Nitzchia* blooms, were more prevalent in areas impacted by river discharges and with weaker upwelling.

An analysis of HAB species distribution in eastern boundary upwelling systems was conducted by Trainer et al. (2010) [73], comparing the distribution of species responsible for HAB blooms in each system and looking for patterns that could indicate particularities of the given system. Four upwelling system were researched in this paper, namely, the California Current, Canary Current, Benguela Current, Humboldt Current. Relative to the Canary Current System, where the Portuguese mainland coast is included, *Gymnodinium catenatum* and *Alexandrium minitum* were found to be the two species most associated with PSP intoxication. The second being more commonly found in more shielded waters, occurring mostly in spring and summer, and being more spatially restricted and associated with stratified conditions. The first one occurs mostly in late summer, is associated with upwelling, appearing earlier on the Portuguese coast than on Galician and being intermittent over the years but having a devastating effect. For DSP, the main outbreaks are associated with *Dinophysis acuminata*, *Dinophysis acuta* and *Dinophysis caudata*, with first two being associated with the most severe HAB, being present during whole year and growing during moderate upwelling. Blooms of *Lingulodinium polyedrum* were reported for some time in the Lisbon region and were associated with low YTX concentrations, water discoloration and bioluminescene. ASP outbreaks and its related Domoic acid (DA) concentrations are mostly linked to the presence of *Pseudo-Nitzchia* species, its presence is recurrent but intensifies normally in the mid upwelling season.

In another study, Fenberg et al. (2015) [19] used a RF model to study how physical environment influenced species distribution across the northeastern Pacific coast, a place where upwelling has a strong impact on the coastline. The RF was used to identify the major biogeographic regions in the study area, through temporal and spatial distribution of algae and macroinvertebrates species combined with twenty-nine environmental variables. The model divided the area in six biogeographic regions that matched previous studies, indicating that a combination of oceanographic and atmospheric variables can

predict the biogeographic structure with high accuracy. The model was also capable of predicting the origin province of almost all samples correctly, and the most important features for species distribution were identified as being nutrient concentrations, SST and upwelling/downwelling season switch index. Some of the regions were found to be more affected by upwelling than others, and on those, UI also revealed a key predictor. A simpler model was then constructed using only the most influential features and obtained a similar accuracy. The results of this study, using algorithms and features discussed for this work in a similar environment, consolidate the premises of this dissertation.

2.3 Relevant Features

Based on the reviewed literature it is possible to identify key features for the study of shellfish contamination and its relation to coastal upwelling. In this section, information about feature use and their combinations in the literature will be compared and discussed, with a focus on the most commonly used features overall, as well as the most commonly used pairs of features.

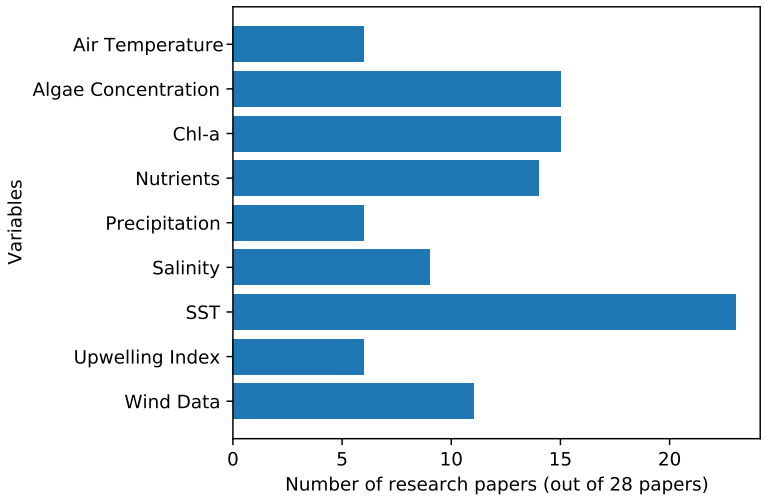


Figure 2.1: Variables occurrence in the literature reviewed in chapter 2.

To analyse the frequency of explored features in the state of the art studies regarding HAB events, shellfish contamination and coastal upwelling, the barchart in Figure 2.1 shows the number of researched papers where the features present on the left side have been explored. As most studies involve more than one feature, the combination of features used throughout literature includes patterns that are important to analyse. A heatmap is displayed in Figure 2.2 containing the proportion between the amount of times the aligned pair of features from the vertical and horizontal axes were used together and the total use of the variable in the vertical axis. Another view of this information is present in the chord diagram in Figure 2.3 where, each feature's perimeter is proportional to the amount of

times it is used in combination with other variables and the width of each arrow between two features varies according to the number of times they are used together.

In Figure 2.1, the most commonly used variables in the literature are displayed, highlighting the frequent use of SST. Other features can also be found numerous times across the reviewed literature including chl-a and wind data. From Figures 2.2 and 2.3 it is possible to observe variables frequently used together, as air temperature and precipitation that are commonly used in combination. Overall, SST maintains its high relevance being frequently used in combination with all the other features displayed. This importance can be easily recognized by the bright column in Figure 2.2 and the wide perimeter and thick arrows in Figure 2.3.

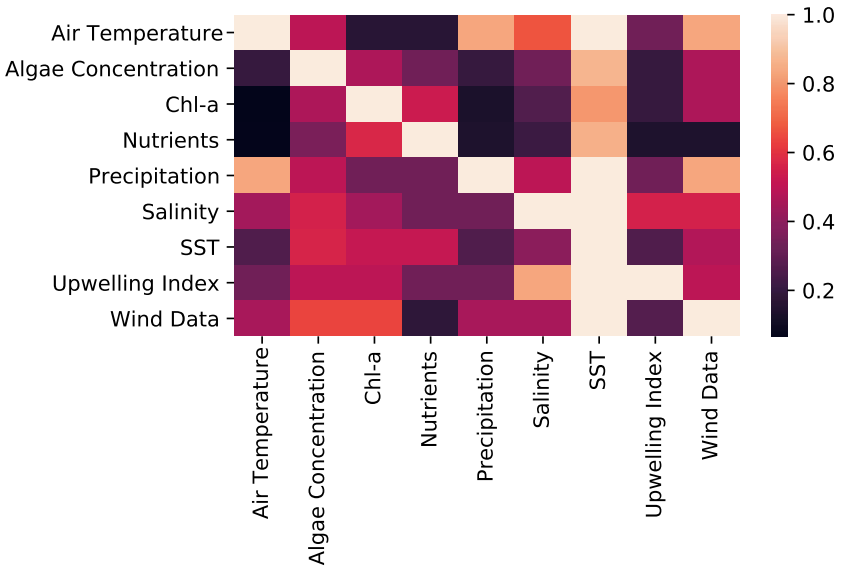


Figure 2.2: Heatmap representing the proportion between the amount of combined uses of each variable pair and the total amount of uses of the variable.

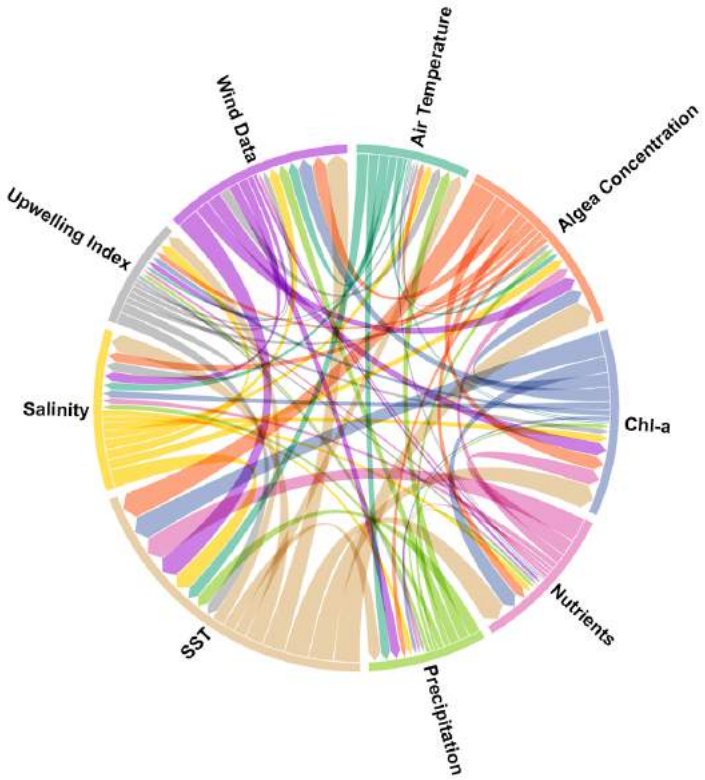


Figure 2.3: Chord diagram displaying the relations between variable combinations.

3

Background Knowledge

The theoretical concepts required for the understanding of the planned work are presented in this chapter. This includes techniques and algorithms considered necessary to produce robust models identified in the previous chapter. Firstly the concept of feature selection and its most common application methods will be discussed, followed by a discussion on supervised learning methods, focusing on SVM and RF, and on evaluation metrics that are commonly used for evaluating the performance of these models.

3.1 Feature Selection

When building a machine learning model, its performance is heavily dependent on the input data, which has to be carefully collected and selected. Feature selection, provides another step to improve the quality of the input, through evaluating performance of features present in the data, in order to use only the features that contribute to the model performance or to reduce dimensionality, therefore reducing the number of features without a significant drop in performance. Feature Selection methods as represented in Figure 3.1 can be categorized in three types: Wrapper, Embedded and Filter methods [60].

- **Filter methods:** The main idea of filter methods is to evaluate each feature on its own, based on statistical measures, independently of the algorithm. With this technique, it becomes possible to choose only the more significant features, in a simple and fast way. The most commonly used filter methods include Chi-squared test, fisher's score, mutual information and correlation.

- **Wrapper methods:** This type of method's approach, on the other hand, involves training the algorithm iteratively with subsets of the set of features, assessing the relevance of each feature based on the model performance when the feature is present in the subset. This technique is able to achieve great results, but with a high computationally cost. The most commonly used wrapper methods include Recursive Feature Elimination (RFE), where the most irrelevant feature is recursively removed until the desired number of features is achieved; Forward Selection, where,

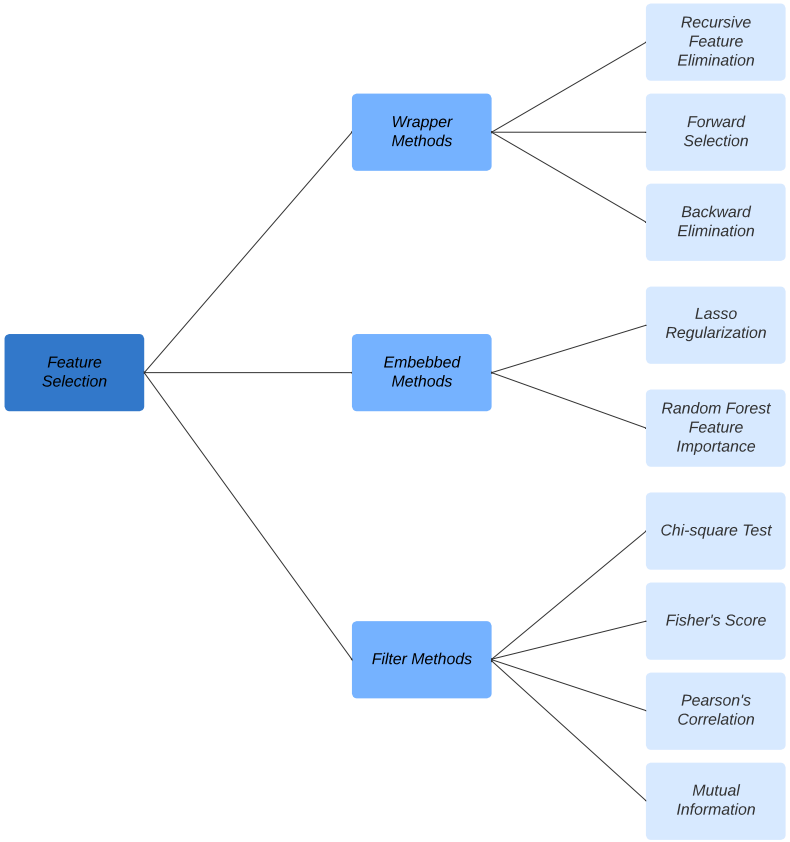


Figure 3.1: Feature Selection Types.

starting with an empty set, features are added one by one until the add of a new features doesn't increase the model performance; and Backward elimination, where, starting with the full set of features, features are removed one at a time until the reduction of features doesn't increase the model performance.

- **Embedded methods:** These methods can be defined as feature selection methods built in the proper model and part of its training phase, evaluating features performance for each iteration during training. The two most common examples of embedded methods are the RF Feature importance and Lasso Regularization.

3.2 Data Split

To create and assess prediction models, a data splitting step is essential. This step includes dividing the available data into training, validation, and testing sets, as represented in Figure 3.2. To make sure that the prediction models are trustworthy and can be applied to new data, it is imperative to complete this phase. In time series analysis, the main goal of data splitting is to train the model on a part of data and assess its performance on a different subset of data that wasn't used for training.

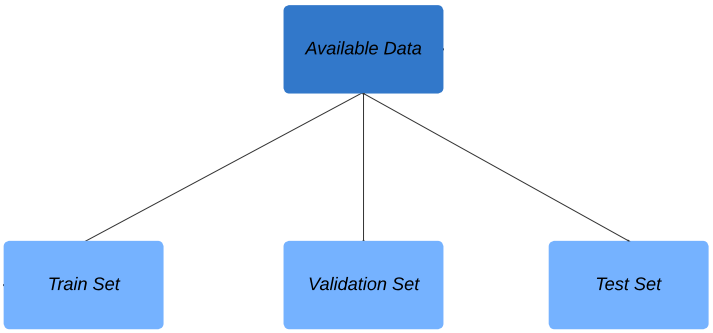


Figure 3.2: Data split general approach.

Data splitting also aids in the detection of overfitting, a problem that frequently arises in time series research when the model does well on the training data but fails to generalize to new data. Time series data differs from standard machine learning problems in that it has a temporal dependency and that the order of observations matters. To prevent data leakage, which can result in exaggerated estimates of model performance, it is crucial to divide the time series data in a way that respects the temporal structure.

3.2.1 Walk Forward Approach

To respect and conserve temporal structure in the data during the data split phase, a walking forward approach can be used. This technique allows for a cross-validation where data is split into sets through a sliding window. In this data split approach, the dataset is split into several folds, with the latest fold, corresponding to the most recent data, being reserved as the test set. The train and validation sets start from the oldest data, as the first and second fold, respectively. In the walking forward cross-validation, the model being constructed is iteratively trained and validated with the previous sets. At each iteration, the train and validation sets move to their next folds until the validation fold strictly precedes the test set. This cross-validation variation is exemplified in Figure 3.3. Two variations of this technique exist, with changes on the train set. In the first one, as represented in the previous figure, the train set moves to the validation fold leaving the previous data behind. In the second variation, the train set maintains the previous folds and, at each iteration, incorporates the new fold.

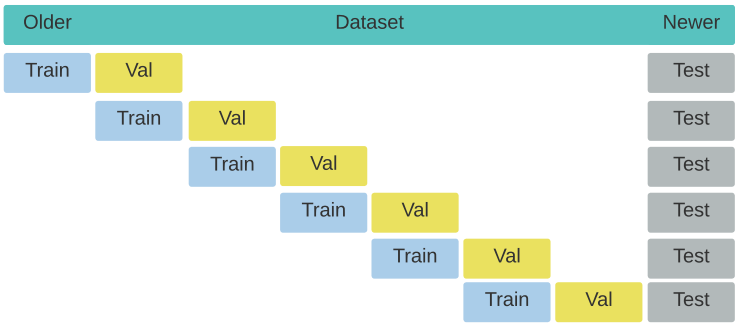


Figure 3.3: Data split example with an Walk Forward approach.

3.3 Supervised Learning

Machine learning methods can be categorized in unsupervised or supervised methods, the latter can also be separated in two techniques: Classification and Regression [6].

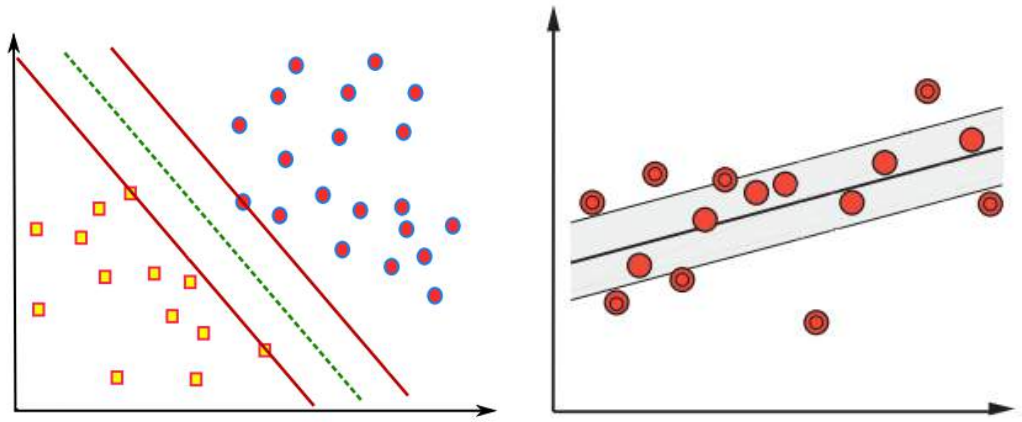
- **Classification Models:** For classification problems the main goal is to correctly predict the label of new sets of data containing features after being trained with a previous set data. This training data contains features and an assigned label. The most commonly used algorithms for classification include naïve Bayes, k-nearest neighbours, and SVM. Classification can be used for problems as blood cell image classification [39].

- **Regression Models:** For regression problems the main goal, on the other hand, is to predict a continuous or discrete numeric value from a set of data. In the same manner as classification, it is trained with a set of features and a corresponding value, instead of a label. The most commonly used algorithms for regression include linear regression, ANN regression, and RF. Regression can be used for tasks as house price prediction [59].

Essentially, Classification and Regression are two techniques of supervised learning that differ in its main objective, with one predicting a label and the other a numeric value as evidenced in Figure 3.4. Another differences between the two also can include the algorithms used for each, as well as, the evaluation metrics used to evaluate the performance and compare between models.

3.3.1 Random Forest

Random Forest [9] is a supervised ensemble learning method that can be used for classification or regression problems. An overview of its steps is represented in Figure 3.5. It is composed of several decision trees that make up the forest, and each tree is built using the bagging technique, in which a random subset with replacement of the full data is



(a) Classification Example. (b) Regression Example.

Figure 3.4: Classification and Regression examples. Adapted from [12, 67].

used, making it random. The use of this resampling technique, also known as bootstrap aggregating, helps to reduce variance by feeding each tree with random subsets of input and averaging the outputs. Diversity between trees can be enhanced using feature sampling techniques within the boostrap sample, that can help reduce correlation between individual trees. As reported by Ziegler et al. (2014) [82], several approaches have been proposed, including using all the features available in the sample, random sampling of features at each splitting node within the tree and a combination of the previous two, keeping some fixed variables and randomly sampling others for each splitting node. To train and validate a RF classifier, generally two thirds of the sample size are used for training, with the remaining being referred as the Out of Bag (OOB) samples, which are then used for validation.

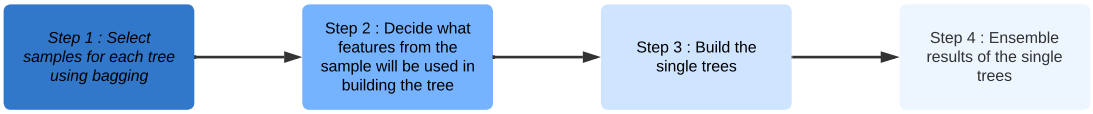


Figure 3.5: General steps of a Random Forest. Adapted from [82].

In a decision tree, the input data is recursively split based on features present in the input and a splitting criterion, terminating when a condition, as maximum depth, is met. For classification problems, common splitting criterion include, Gini impurity, that measures the probability of new data being misclassified based on the distribution of the classes and Entropy, that measures the reduction in uncertainty resulting from the chosen splitting condition. In the case of regression, weighted variance can be used as the splitting criterion. The combination of using multiple trees instead of a single

one, makes the model more robust and increases its generalization power. The results of the RF are obtained from ensembling the results of the single trees, in the form of a majority vote for classification or as an average for regression, making the final result more robust and accurate. RF are commonly built using Classification and Regression Tree (CART) decision trees, but other variations as Modified Classification and Regression Tree (mCART) and other types can be used. Although reports suggest that RF can achieve an high performance with default parameters, these can be tuned to further improve performance [62]. These parameters include, the number of Decision Trees used, that balances the tradeoff between performance and computational cost, the node size, that controls the minimum number of observations in a leaf node, mtry, that regulates the number of variables used when selecting the next split and sample size, that determines the size of the sample used in each tree.

The main advantages of using RF include its increase in accuracy compared to single trees, its capacity of handling large sets of data with high-dimensionality and non-linear relationships among variables, while not being very sensitive to noisy data and missing values and containing a built-in feature selection method. On the other hand, RF does not allow for a detailed interpretation of its behaviour and requires a significant amount of time and resources for the training phase. RF models can be used in several domains of application such as finance and journalism [68]. In the finance area, machine learning models can be used to evaluate a client's profile and assess risk on requested operations. RF models have been used to forecast the probability of clients defaulting on their debts based on information about them and their past billing records. In the journalism area, users' interests can be tracked and targeted by custom advertisements or suggestions. Several RF models have been developed to forecast the popularity of news articles based on several aspects of the article's content and its publishing.

3.3.1.1 Iterative Random Forests

IRF is a variation of the RF algorithm. The main objective of this variation is to improve the overall accuracy of the model by iteratively enlarging the forest. In an IRF, for each iteration a new decision tree is added, while maintaining the previous trees and its previous predictions, which are taken into account to make the predictions more accurate. This leads to a more robust model that can generalize better to new data. In contrast to a standard RF, in an IRF the same data is used for all decision trees in the ensemble, with a random subsample of features being used as the input for each tree. IRF also differs from standard RF in the way the predictions are combined, instead of a majority vote or a standard average, a weighted average of the single trees output is used, where the weights represent the tree accuracy.

3.3.1.2 Quantile Random Forests

QRF is a variation of the standard RF algorithm that can be used for regression problems. In contrast to a standard RF, instead of a majority vote, this algorithm returns several predictions, each one for a different quantile of the output. This variation is possible through training several trees for each quantile and ensembling the outputs as a prediction for the given quantile. The quantile specific predictions allows not only for a better customization and risk assessment in forecasting problems, but also for a better understanding of output distribution.

3.3.1.3 Feature Selection using Random Forest

Algorithms like RF and its variations have a built-in feature selection method, known as feature importance. The main goal of **feature importance**, as other methods, is to measure the impact of each features in the global performance of the model, and determine the degree of importance of each feature. The relevance of each feature is calculated based on the decrease in impurity when used to split the data, using metrics as Gini Importance and Mean Decrease Accuracy (MDA). In contrast to being conveniently built-in in RF, feature importance possesses some downsides as being extremely dependent on the model robustness and showing a bias towards features with high cardinality.

3.3.2 Support Vector Machine

Support Vector Machine [15] is a supervised machine learning algorithm typically used for classification problems. The main goal of an SVM is to separate the input into different classes, finding the best boundary or hyperplane. The general steps required to construct a SVM model are represented in Figure 3.6. To accommodate regression problems, the goal of the algorithm is switched to finding the hyperplane that fits the input the best, with this variation being known as a Support Vector Regression.

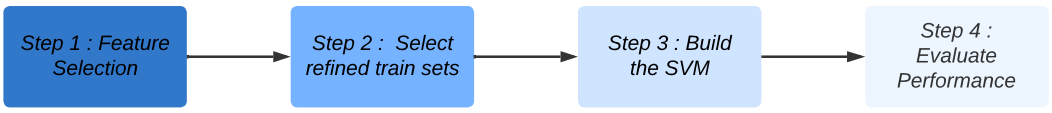


Figure 3.6: General steps to build a Support Vector Machine.

After an initial stage of feature selection, the model performance can be further improved through the use of algorithms capable of selecting refined training sets from the dataset [50]. Several techniques are available, offering different optimization strategies, represented in Figure 3.7, which include random sampling methods, where a train set is built from random samples of the dataset; data geometry analysis methods, where samples deemed irrelevant are removed, selecting the samples closer to the hyperplane or

in heterogeneity areas, considered more important; neighbourhood analysis methods, try to select correctly labelled samples near the hyperplane, that are considered to be more impactful. These methods are extremely important when dealing with large datasets as they increase the speed of the model, reducing the amount of data and removing outliers that could hinder the performance.

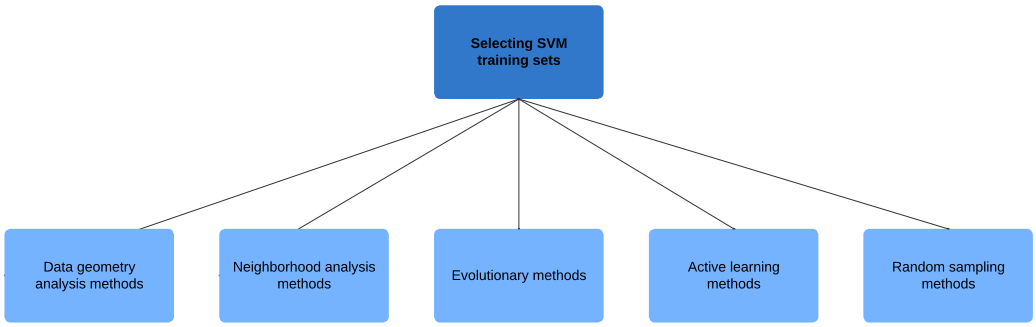


Figure 3.7: Methods to select refined train sets for SVM. Adapted from [50].

This algorithm starts by plotting the input data in a dimensional place sized accordingly to the number of features and finding the closest points from each class to the hyperplane, with these becoming known as support vectors. The final step is to choose the best suited hyperplane based on the maximum margin between it and the support vectors.

SVM models can be linear or non-linear according to the hyperplane used to separate the data. In a linear SVM, a linear hyperplane is used to separate the data classes. For non-linearly separable datasets, non-linear SVM can be used, where a non-linear hyperplane is generated to separate the classes. These datasets can also be mapped into a higher dimensional space where they can be linearly separated through the use of the kernel trick. For slightly non-linearly separable datasets or to increase the maximum margins, soft margins, which allow some misclassifications, can be used instead of hard margins. Soft margins introduce a regularization parameter, that penalizes wrong classification.

For the model success, two important parameters have to be properly configured, namely, the kernel function and the regularization parameter. The data plot referred previously, as the first step in this algorithm, is carried out by the kernel function. In this manner, it is important to select the right kernel function accordingly to the characteristics of the data. The most commonly used kernel functions include linear, polynomial and RBF kernels. The regularization parameter represents the degree of misclassification errors accepted by the model, through a tradeoff between margin size and misclassification error.

The main advantages of using SVM, is it capacity to work in high-dimensional spaces, not requiring a large dataset to obtain good performance and being able to handle non-linear relations. On the other hand, it is computationally expensive, taking more time and

memory than other algorithms, requires a long training phase, is heavily dependent on proper configuration of kernel function and regulation parameter and being by default limited to binary classification. SVM models can be used in several domains of application such as botany, bioinformatics, energy production and hand-writing detection [67, 12]. In the botany area, SVM models have been used to classify plant species from pictures. SVM models have also been widely used in the bioinformatics field, with models developed for problems as cancer classification, using data as cell images or gene expression, or cardiovascular organs classification through texture information. In the energy production area, energy sources as wind have a high temporal intensity fluctuation, making its distribution unreliable. To tackle this problem, several Support Vector Regression (SVR) models have been developed to forecast wind speed based on previous wind speed records. SVM models have also been in hand-writing recognition, classifying characters based on identified features and patterns.

3.3.3 Evaluation Metrics

In order to evaluate the performance and compare models, evaluation metrics have to be used. Evaluation metrics have an important role and have to be carefully selected according to the technique being used. When dealing with a classification problem, the most commonly used evaluation metrics are accuracy, precision, recall, F1-Score and Receiver Operating Characteristics (ROC) and Area Under The Curve (AUC), if dealing with binary classification [6].

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ positives + false\ negatives}, \tag{3.1}$$

$$precision = \frac{true\ positives}{true\ positive + false\ positives}, \tag{3.2}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}, \tag{3.3}$$

$$F1\ Score = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{3.4}$$

The accuracy metric given by Equation 3.1 represents the proportion of correct predictions made by the model, this can, however, be a misleading metric when dealing with unbalanced classes distribution, as it can give a high score even if the model is not able to correctly classify the least frequently occurring class. The precision metric given by Equation 3.2 represents the proportion of true positive predictions among all positive predictions. The recall metric, also known as True Positive Rate (TPR) and Sensivity, given by Equation 3.3 represents the proportion of true positive predictions among all actual positive instances. The F1-score metric given by Equation 3.4 represents the harmonic mean between the precision and recall metrics. The ROC curve is generated by plotting the false positive rate against the true positive rate at different classification thresholds. This metric helps find a desirable threshold by analysing the tradeoff between the two rates. The AUC provides a measure of the overall performance of a classifier in the classification thresholds.

Variations of the previous metrics and other metrics can be better suited when evaluating a classification problem with class imbalance. These include the average precision, balanced accuracy, F1 Macro and the Precision-Recall (PR) Curve and its AUC.

$$average\ precision = \sum_{i=1}^{n}(precision_i \cdot \Delta recall_i), \tag{3.5}$$

$$specificity = \frac{true\ negatives}{true\ negatives + false\ positives}, \tag{3.6}$$

$$balanced\ accuracy = \frac{sensitivity + specificity}{2}, \tag{3.7}$$

$$F1\ Macro = \frac{1}{n} \cdot \sum_{i=1}^{n} F1\ Score_i. \tag{3.8}$$

The average precision, given by Equation 3.5, represents a weighted mean of the precision scores obtained for every i threshold in n^{th} thresholds with the variation in recall from the $i-1$ threshold being used as the weight. The balanced accuracy, given by Equation 3.7, represents the mean between the sensitivity, recall or TPR, given in Equation 3.3, and the specificity or True Negative Rate (TNR), given in Equation 3.6. The F1 Macro metric, given by Equation 3.8 where n represents the number of distinct classes, considers the F1-Score for each class individually, attributing the same weight to each class through adding all the F1 Scores and dividing them by the number of classes. The PR curve presents the precision and recall scores at several classification thresholds, allowing the analysis of the trade-off between the decrease in precision and the increase in recall. Similarly to the ROC curve, its AUC, helps summarize the overall performance of the classifier.

In contrast, for regression problems, the most commonly used evaluation metrics are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2) and Mean Absolute Error (MAE), which can be given by :

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{3.9}$$

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.10}$$

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{3.11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.12}$$

where y_i represents the actual value and $\hat{y}_i$ the predicted value for every observation i in the data n and $\bar{y}$ represents the mean value of the y_i values. MSE metric represents the average of the squared difference between predicted and actual values. MAE metric, similarly, represent the average of the difference between the predicted and actual values without them being squared, unlike MSE, making it insensitive to outliers. R^2 metric

represents the degree of variation of a dependent variable that can be explained by an independent variable.

4

Experimental Setup

In this chapter, the proposed methodology is initially presented followed by a detailed description of the dataset construction process. A data analysis is then presented for the constructed datasets and a walkthrough of the main steps of the models' construction. Finally, the hyperparameter tuning phase of the models is analysed and discussed. The main objectives of the designed experimental setup are the following:

- Merge the different data sources, preprocess their variables and build the datasets.
- Analyse the datasets, understand its variables and assess their contamination class imbalance.
- Select the appropriate sliding window and hyperparameter values' range and evaluation metrics.
- Construct and compare models following classification and regression approaches.
- Apply and analyse the different algorithms for each studied approach; RF and SVM for classification; RF and SVR for regression.
- Evaluate and compare the performance of the models when using different sets of variables.
- Assess the predictive role of environmental variables on shellfish biotoxin contamination.

4.1 Proposed Methodology

To achieve the best results, a robust experimental protocol was developed. This protocol tackled the problem as a classification and a regression task, building distinct datasets accordingly. These datasets included different outcome variables, a binary variable for contamination status for classification and a numeric value encoding the level of contamination for regression. The data used in this work was structured as several weekly

time-series and was obtained from previous works and partner institutions. These time-series included several features such as: toxin concentration in shellfish, mean, minimum and maximum air temperature, mean wind direction, mean wind intensity, precipitation intensity, salinity, water level and total upwelling area. Due to the time component present in the data, the datasets were converted into a sliding window format and split using a walk forward approach, described in subsection 3.2.1.

The constructed models were built using different algorithms, comparing the performance between them. The chosen algorithms included RF and SVM for classification and RF and SVR for regression, that although seldomly used in the reviewed literature for the studied problem, received increasing attention in recent studies and their potential was further explored in this work. Features were selected based on their availability, completeness and use throughout the reviewed literature, as evaluated in Section 2.3. Finally, the models were evaluated using the proper evaluation metrics identified in section 3.3.3. From the obtained results, the performance between algorithms and approaches was compared, leading to conclusions on the most favorable strategy for the problem and evaluating the importance and contribution of several environmental variables on shellfish contamination. A rundown of this process is presented in the flowchart in Figure 4.1 with the main aspects summarized below.

- **Problem formulation:** Developed classification and regression approaches for the problem using distinct datasets with different outcome variables, a binary variable for contamination status for classification and a numeric level of contamination for regression.
- **Dataset construction:** Built the datasets using data collected from two previous master thesis [17, 44] and partner institutions.
- **Data structure:** Performed data split on the acquired time-series using an walk forward approach to generate train, validation and test sets.
- **Models' construction:** Selected the most important features in the datasets and built the models using the selected algorithms (RF and SVM) and evaluated the models' performance using the appropriate evaluation metrics.
- **Results evaluation:** Analyzed the obtained results and compared the algorithms and approaches, to identify the best strategy for the problem and evaluated the influence of different environmental variables on shellfish contamination.

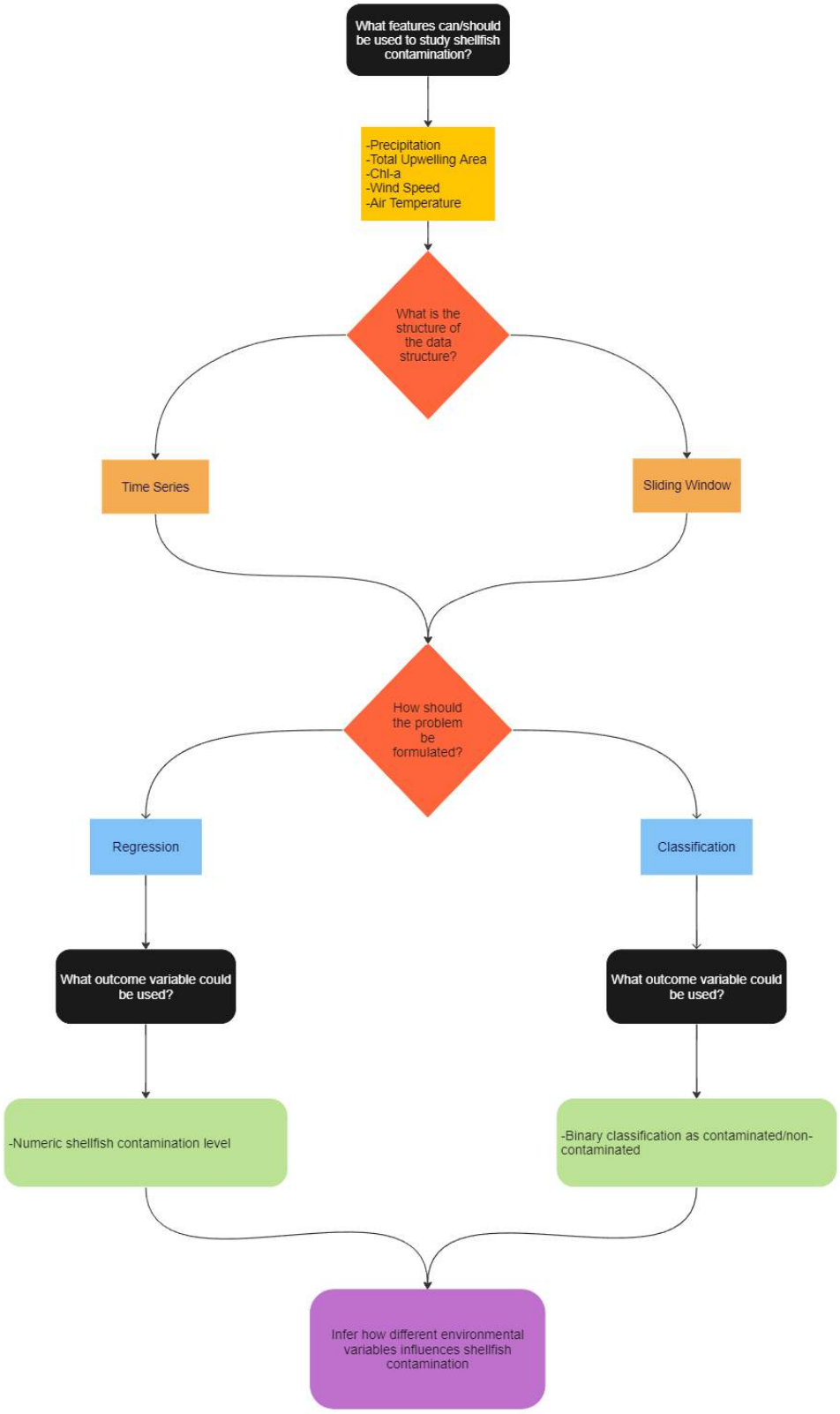


Figure 4.1: Summary of the steps for the models' development.

4.2 Dataset Construction

From the obtained data, several datasets were constructed, taking into consideration the specific requirements of each approach. These datasets are composed by different quantities and combinations of features. These variations allow the evaluation and comparison of the importance of distinct groups of features. In order to improve the quality of the data and its completeness, as time interval regularisation and missing values interpolation were used during the construction process.

As a basis for all datasets, datasets composed only of DSP data were created for every available production region as described in Subsection 4.2.1. Following this step, the enrichment of the basic datasets was performed by adding complimentary features through processes described in subsubsection 4.2.2 for meteorological variables, subsubsection 4.2.3 for HWP variables and subsubsection 4.2.4 for coastal upwelling related variables.

4.2.1 DSP Concentration

The DSP concentration data was collected by IPMA for the period between 2014 and early 2023 through regular sampling periods on a weekly or biweekly basis in every shellfish production region, displayed in Figure 4.2.

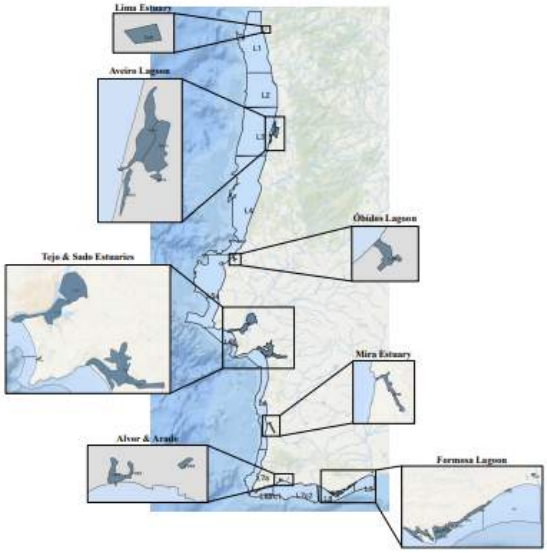


Figure 4.2: Shellfish Producing Regions. Taken from [17].

In order to achieve the desired DSP data format to create the initial datasets, several steps were taken.

- The biotoxin samples were sorted by sampling date.
- The samples with registered DSP concentrations equal to 0.0 were replaced by the minimum detectable value of the used technique, 28.0.

- The data was then split by production zone and the samples from different assigned sampling regions within the zone were separated. An example of this separation can be seen in Figure 4.3 where the data of the L1 zone is separated into L1 Carreço and L1 Labruge with misregistered observations being removed.

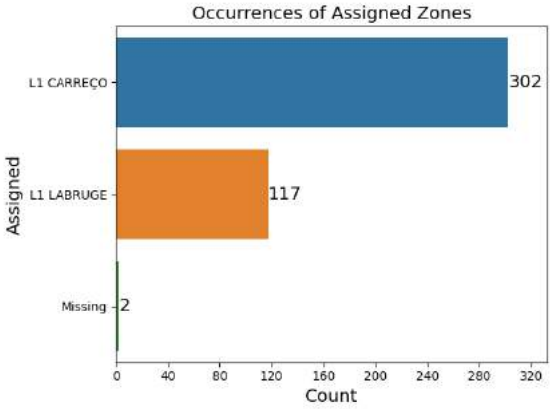
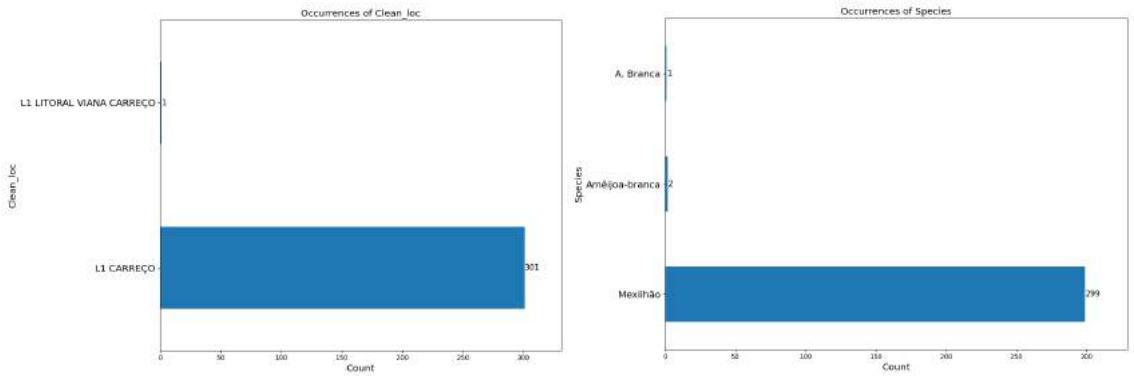


Figure 4.3: Occurrences in each assigned sampling zone in the L1 region.

- From within each assigned sampling region the actual sampling location was verified to identify possible errors when registering the observations. As an example the actual sampling location on the L1 Carreço assigned sampling zone are displayed in Figure 4.4(a).

- Possible errors when registering the sample's specie were investigated and the DSP concentration values of the predominant sampled specie were used to create the dataset. This step is exemplified in Figure 4.4(b) where A. Branca and Amêijoa-Branca are actually the same species and Mexilhão is selected due to being the most predominant sampled specie.



(a) Clean location distribution.

(b) Species distribution.

Figure 4.4: L1 Carreço clean location and species distributions.

- Due to irregular sampling intervals, the week of every sample was calculated using **datetime's isocalendar().week** and its registered sampling date in order to ensure weekly observations. Since there are samples from several years, to distinguish between same weeks of distinct years, the difference between each sample's year and the first registered year times the number of weeks in a year is added to every value, making it unique. This calculation is highlighted in Listing 4.1.

Listing 4.1: Sample Week

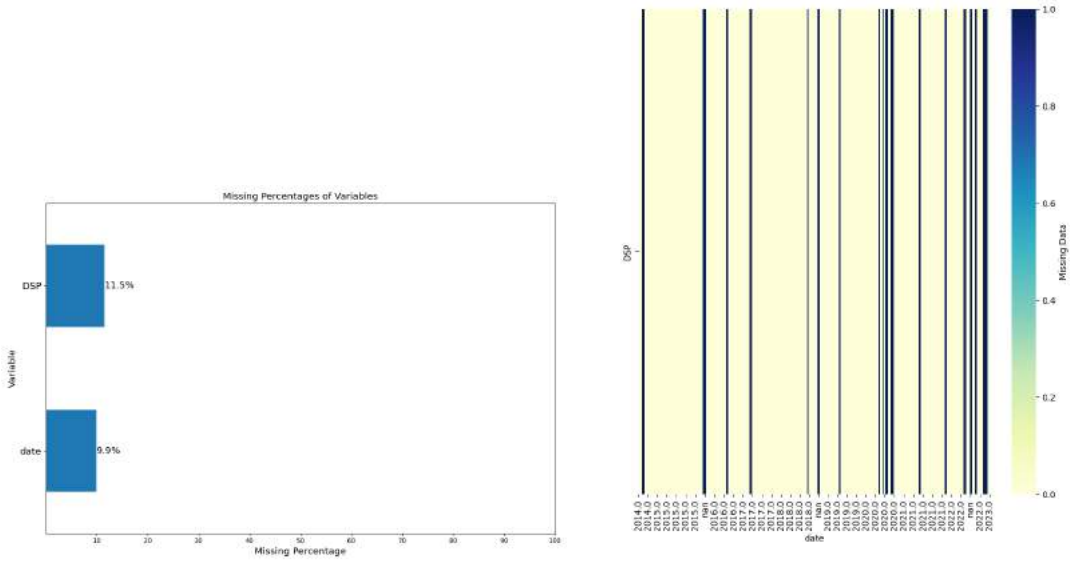
```
import datetime

sample_week = sample_date.isocalendar().week + (sample_year -
    firstsample_year) * 52
```

- As ensuring a weekly format on the dataset revealed several instances where the interval between consecutive samples slightly exceeded a week, weeks with no registered values were added. If a week with no registered values precedes and is followed by valid weeks its value is calculated as the average of the previous and following weeks.

- In order to select the best datasets, the percentage of missing values and their temporal distribution was examined, as exemplified in Figure 4.5. These results allowed the identification of the zones with the lowest missing values percentages and the detection of time periods where these missing values could be concentrated.

To compare the results between all zones and to select the best ones in an convenient way, a scatter plot containing all zones, their missing values percentage and total number of observations was constructed.

In Figure 4.6, the regions present in the top left corner contain an overall higher number of samples and an overall lower percentage of missing values. These regions have the most complete data and will be further analysed in Section 4.3.



(a) Missing percentages. (b) Missing periods.

Figure 4.5: L1 Carreço DSP missing values.

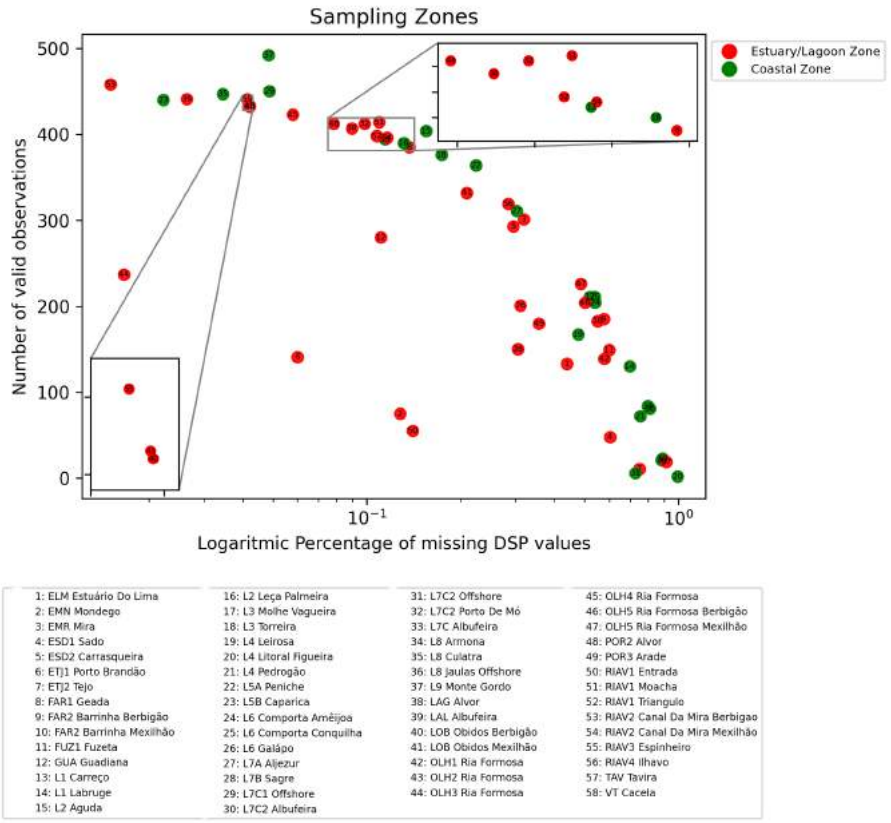


Figure 4.6: Scatter plot for the missing percentage and number of observations.

4.2.2 Meteorological Variables

Similarly to the DSP concentration, the meteorological data was collected hourly throughout the year by IPMA during the same time period by the closest weather station to every shellfish production region. Introducing this data to the datasets allows the evaluation of the impact these variables have on shellfish contamination. The meteorological data contains the variables described in Table A.2. Due to the large number of variables, these were filtered considering the most popular variables identified in reviewed literature. Based on the relevant features described in Section 2.3, only the variables T_MAX, T_MIN, T_MED, FF_MAX, FF_MED, DD_PRE, DD_MED, PR_QTD and PR_DUR were considered. Unfortunately, the meteorological was not fully available for regions as the L8 Culatra.

In order to make hourly variables compatible with the weekly DSP variable, transforming the meteorological variables into weekly variables was necessary. This process consisted of, in a first stage converting the daily 24 measurements into a single daily value. This daily value was calculated considering the characteristics of each variable.

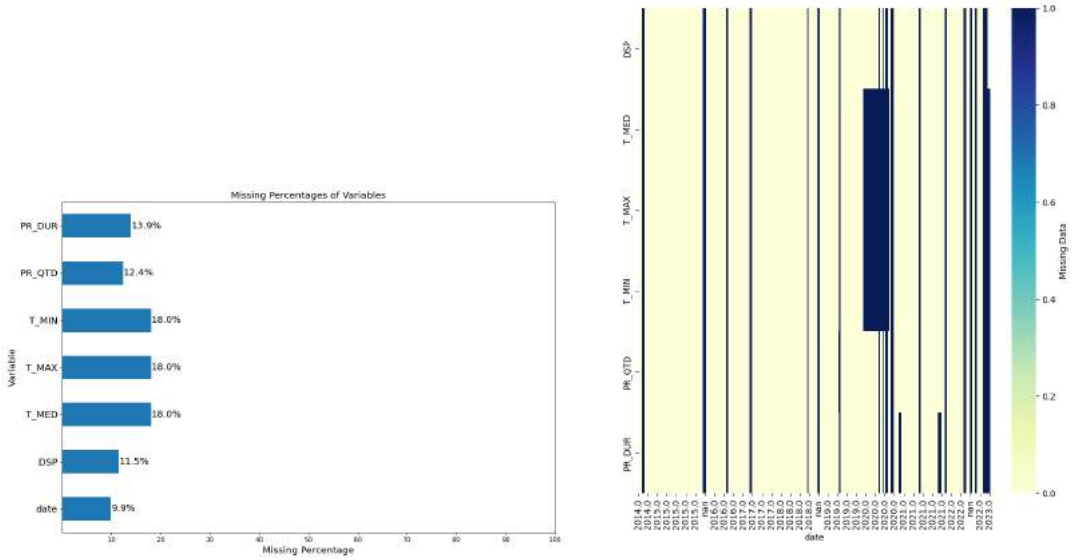
- For T_MED, FF_MED, PR_QTD and PR_DUR the average daily values were calculated.
- For T_MAX and FF_MAX the maximum daily value was calculated.
- For T_MIN the minimum daily value was calculated.
- For DD_MED, due to being in degrees, the daily value was calculated using a circular mean.
- For DD_PRE, as this variable is a categorical variable, the daily value was calculated using the mode.

For every DSP sample, the meteorological daily values, if available, were gathered for the seven days prior to the biotoxin sampling date. These values were converted into weekly values, allowing them to be merged with the DSP data. These conversions followed a similar process to the daily conversions.

- For T_MED, FF_MED, PR_QTD, PR_DUR, T_MIN, T_MAX and FF_MAX the average weekly values were calculated.
- For DD_MED, the weekly value was calculated using a circular mean.
- For DD_PRE dummy variables, the weekly value was calculated using the mode. From the weekly value, ten dummy variables were created, one for each possible category. The weekly dummy variable corresponding to the weekly mode would assume the value of 1 with the remaining assuming 0.

Following the addition of the meteorological variables to the datasets, an analysis on the missing values of these variables was needed in order to ensure the completeness of

the data. Considering that most meteorological variables present an elevated percentage of missing values, from the previously listed variables, only the ones containing less than 20% of missing values were considered, varying across zones. In Figure 4.7 an example of this analysis is presented for the L1 Carreço zone. Due to the inability of the SVM and SVR algorithms to support missing values, the remaining missing values for the environmental variables, including meteorological, were replaced by the 5 month average, centered from the sampling date, or the global average of the value, if the previous average proved impossible to calculate.



(a) Missing percentages.

(b) Missing periods.

Figure 4.7: L1 Carreço meteorological missing values.

4.2.3 Hydrodynamic and Water Properties Variables

In an effort to further enrich the datasets, HWP data was also taken into consideration. This data was collected hourly throughout the year for the same time period as the previously mentioned data, through estimation of several biotic and abiotic variables in the water column with the MOHID-PCOMS system [43, 57, 45, 46, 10] on the DSP sampling coordinates by MARETEC, a partner institution of project MATISSE. This data was only available for a few of the regions being considered, namely L1 Carreço, L9 Monte Gordo and RIAV1 Triângulo and contained the variables described in Tables 4.1 and 4.2. Due to the reduced number of variables, all were taken into consideration and additional variables were created from the maximum and minimum values of the water_level and phytoplankton variables.

Table 4.1: Hydrodynamic Variables

Code	Name	Unit
velocity_V	Velocity towards North	$m^3 s^{-1}$
velocity_U	Velocity towards East	$m^3 s^{-1}$
velocity_modulus	Velocity modulus	$m^3 s^{-1}$
water_level	Water Level	m

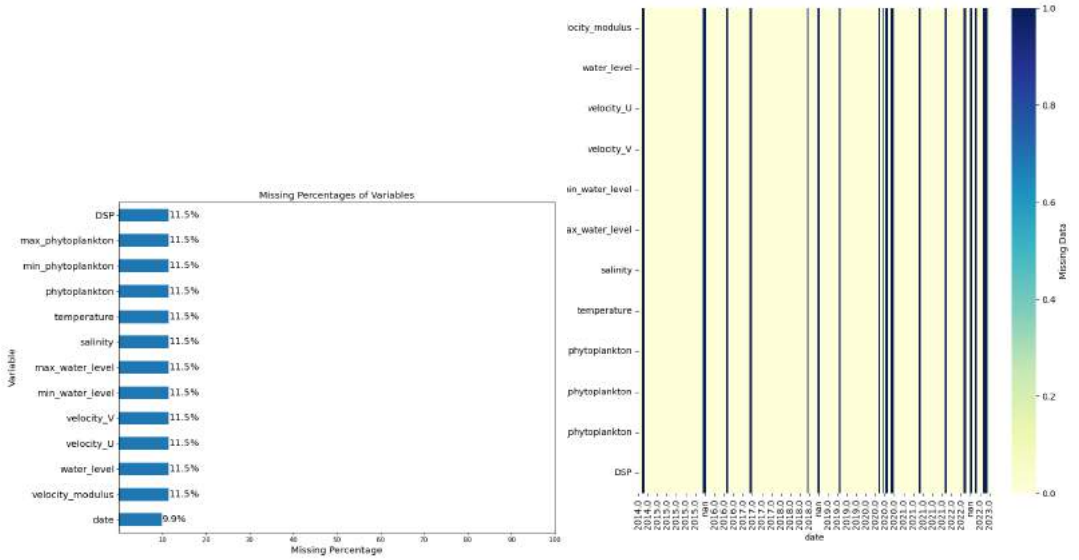
Table 4.2: Water Properties Variables

Code	Name	Unit
salinity	Salinity	ppt
temperature	Water temperature	^oC
phytoplankton	Phytoplankton concentration	$mgCl^{-1}$

Similarly to the transformation process applied to the meteorological data, the HWP variables were firstly converted into daily values according to their characteristics.

- For salinity, temperature, phytoplankton, and water_level the average daily values were calculated.
- For max_water_level and max_phytoplankton the maximum daily values were calculated.
- For min_water_level and min_phytoplankton the minimum daily values were calculated.
- For velocity_V, velocity_U and velocity_modulus the median daily values were calculated.

Following the daily conversion, the variables were averaged into weekly values using the seven daily values prior to all DSP samples, in a process similar to the one used in the meteorological data transformation. Finally, as proceeded for the previous variables, the missing values on these variables were examined and interpolated when needed. This analysis revealed almost no missing values, with the identified ones matching the missing values already present in the DSP data, as exemplified for the L1 Carreço region in Figure 4.8.



(a) Missing percentages.

(b) Missing periods.

Figure 4.8: L1 Carreço hydrodynamic and water properties missing values.

4.2.4 Upwelling Variables

In recent work regarding the automatic recognition of coastal upwelling and its spatio-temporal analysis [55, 54], an automated framework core-shell clustering for the automatic recognition of coastal upwelling from SST data derived from SST satellite data, represented in Figure 4.9, was developed.

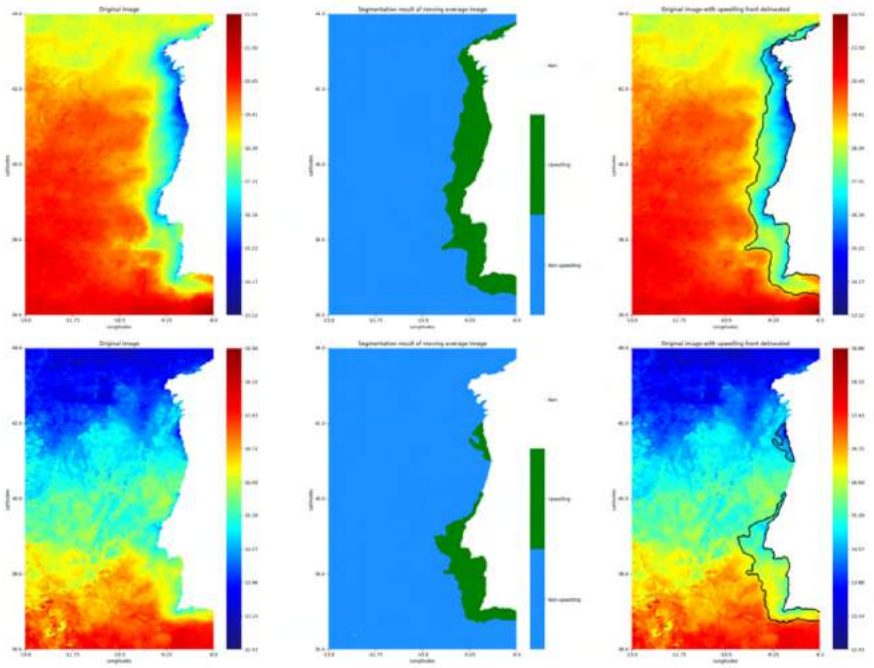


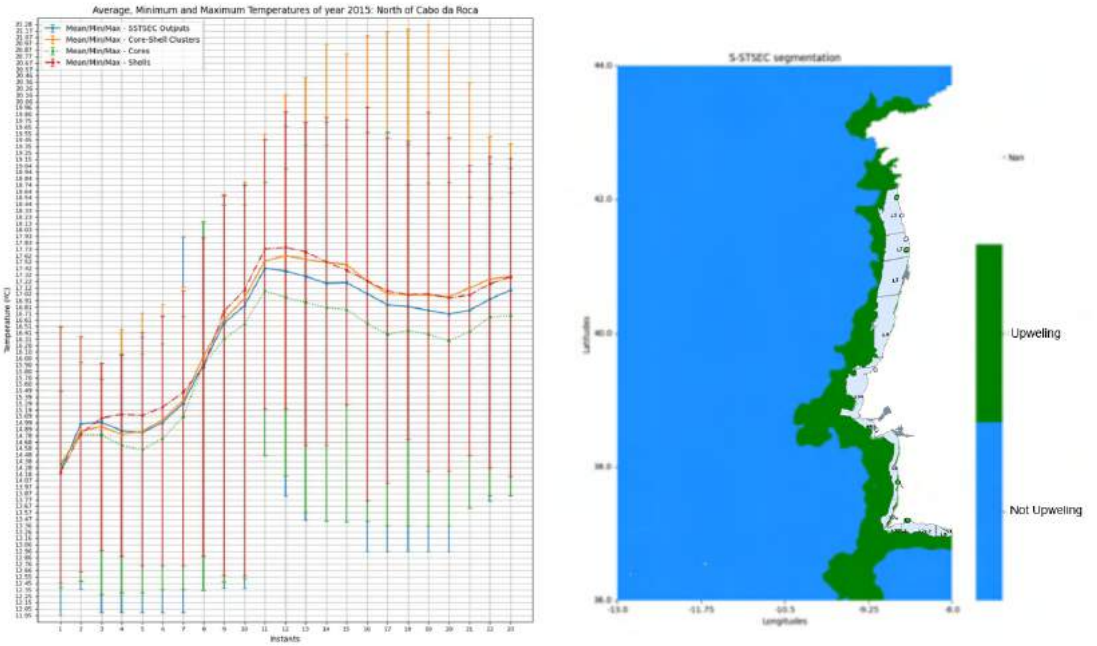
Figure 4.9: SST satellite, core-shell upwelling zone and perimeter. Taken from [44].

The framework incorporates two recently proposed algorithms for unsupervised spatial clustering [52] and spatio-temporal clustering [53]. These algorithms provide hyperparameters automatically derived from the clustering criteria and do not need to be empirically fine-tuned as popular spatio/spatio-temporal clustering (e.g. Density-based spatial clustering of applications with noise (DBSCAN)).

The core-shell clustering was applied to sixteen years of daily SST images acquired from the OceanColor site [51] in the period between 2004 and 2019. These images were averaged from eight daily images into one week images. Since the main goal of this work was the automatic recognition of coastal upwelling and its spatio-temporal analysis, only the weekly images from March 30th to October 30th of every available year were considered.

The coastal upwelling variables and corresponding time-series were extracted from this framework for all assigned sampling coordinates using a radius of 6km and 10km. This data included the variables present in Table 4.3, namely the mean SST, represented in Figure 4.10(a) and the upwelling core-shell region, which can be seen overlapped with the shellfish production areas in Figure 4.10(b). To work with this data, which focuses on the upwelling season, instead of converting it to match the previously described data, it was necessary to convert the other data to the same format.

This transformation was achieved by replicating the previous data's preprocessing on the required data. This included selecting the 27 weeks between the same time period, from the 30th of March to the 30th of October of every year, using 8-days weeks. From the 27 weeks, a 5-week sliding window was used to average the results into the desired 23



(a) Mean SST for each instance.

(b) Core-Shell instance and Shellfish Production Areas.

Figure 4.10: Example instances on Mean SST and Coastal Upwelling Core-Shell Regions. Taken and adapted from [44].

yearly instances format.

Table 4.3: Coastal Upwelling Variables

Code	Name	Unit
average temp	Average temperature in the upwelling region	^oC
total area	Total upwelling area	km
min temp	Minimum temperature in the upwelling region	^oC
max temp	Maximum temperature in the upwelling region	^oC
mean dist	Mean distance to the coast from the upwelling border	km
max dist	Maximum distance to the coast from the upwelling border	km
temp diff	Temperature difference from the upwelling region to the neighbouring region	^oC

The missing values' prevalence was also examined for these variables, as performed for the previous ones, and their values interpolated when needed. Similarly to the HWP missing values results, the coastal upwelling variables presented almost no missing values that were not already present in the DSP variable. Due to problems on the original data it was not possible to obtain coastal upwelling data for the L8 Culatra, L9 Monte Gordo and L7c2 Porto de Mós regions.

4.2.5 Dataset Designation

Due to the large number of datasets created and used in the developed models these will be referred to using the following notation. Firstly, they will be identified by their

sampling region code, followed by an upwelling code, UP, noting that the dataset is in the coastal upwelling data format. An omission of this code indicates that the dataset does not focus on upwelling. Finally, the types of variables used in the dataset are represented by their initial letter; D for DSP; M for Meteorological; H for HWP; U for upwelling variables. Using this designations, for example, the L1 Carreço dataset using DSP and meteorological variables is referred to as L1-DM and the L2 Leça da Palmeira upwelling dataset for DSP and HWP is referred to as L2-UP-DH. A list summarizing the constructed datasets as well as the corresponding designations can be consulted in Table A.1.

4.3 Data Analysis

After producing the datasets it is necessary to study and understand the variables in them. This analysis will be focused on datasets for the zones identified in Figure 4.6, as the most complete data-wise. As a starting point for the analysis, the main variable of this study, DSP concentration, displayed in Figure 4.11, for the L1 Carreço production zone from August 2014 to March 2023 was studied. In Figure 4.11, the dotted red line, indicates the DSP contamination threshold, referred in Table 1.1, allowing the identification of yearly contamination spikes. It is important to consider that this region presents a higher prevalence of contamination events than other studied regions.

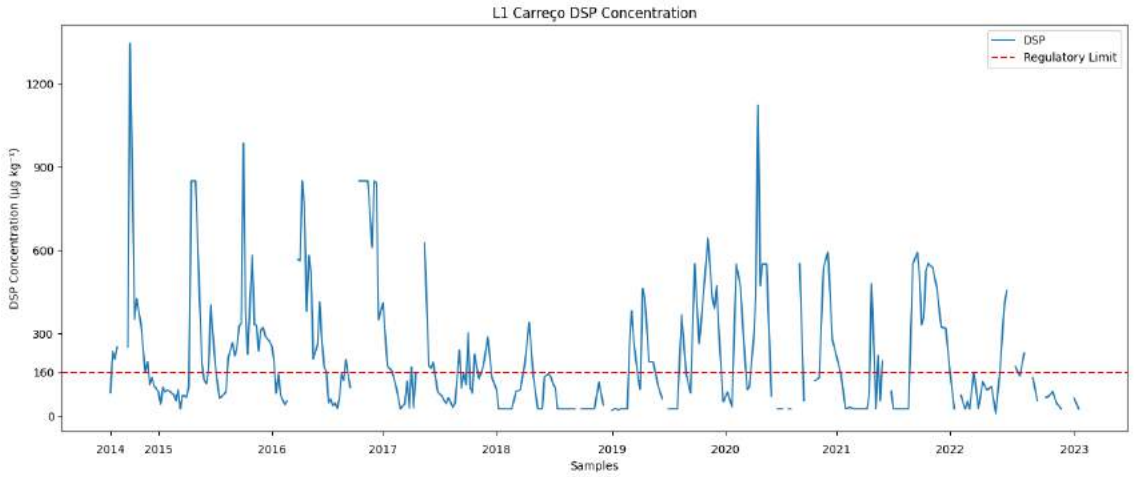
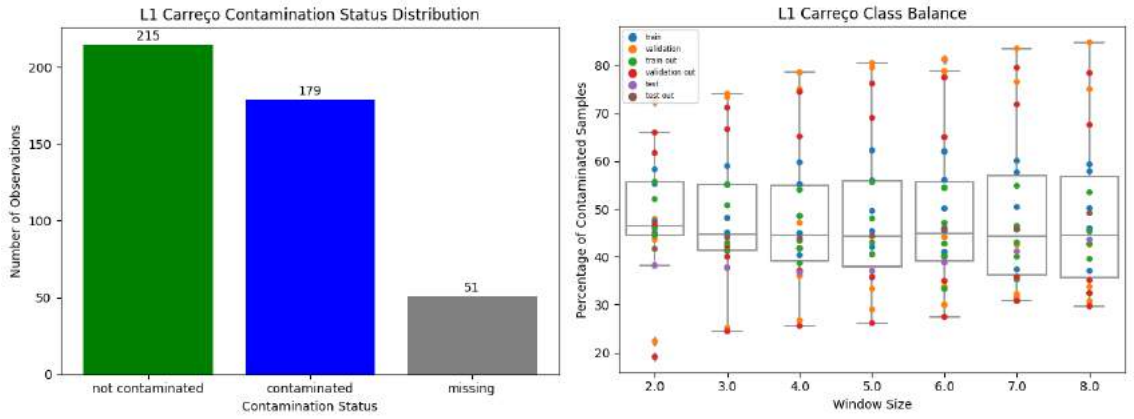


Figure 4.11: L1 Carreço DSP Concentration.

As the main goal of the developed models is to predict DSP contamination it is important to study class balance between contaminated and not contaminated samples, in order to select the best zones, containing a higher frequency of contamination events, and the appropriate methodology for model construction. The methodology used in the models construction requires a careful consideration of the class distribution when converting the data into a sliding window, splitting the data into sets and choosing the appropriate evaluation metrics, processes which are described in detail in Subsection 4.4. Class imbalance is natural when studying a sporadic seasonal phenomenon, characterized

by occasional spikes throughout the year, like shellfish contamination. From Figure 4.12(a), it is possible to identify a high number of contaminated samples, in accordance with Figure 4.11, showing a moderate balance between classes and a low quantity of missing values. As the data will not be used as a whole and in a time-series format it is necessary to study the class balance at the sliding window format and set-wise. In Figure 4.12(b) an example of this analysis can be seen. In the box plots it is possible to observe the percentage of contaminated samples in each set, with a value of 100 indicating that 100% of the samples are contaminated. The medians of the box plots indicate an overall balance across all window sizes. It is also important to note that bigger window sizes present lower averages but a higher variation throughout their sets. For all window sizes, the sets presenting higher imbalances are validations sets, likely the same ones throughout the different window sizes, representing time periods either dominated by contamination events or with few contamination events. This can be justified by the varying sizes of the sets that are required to follow the natural timeline.

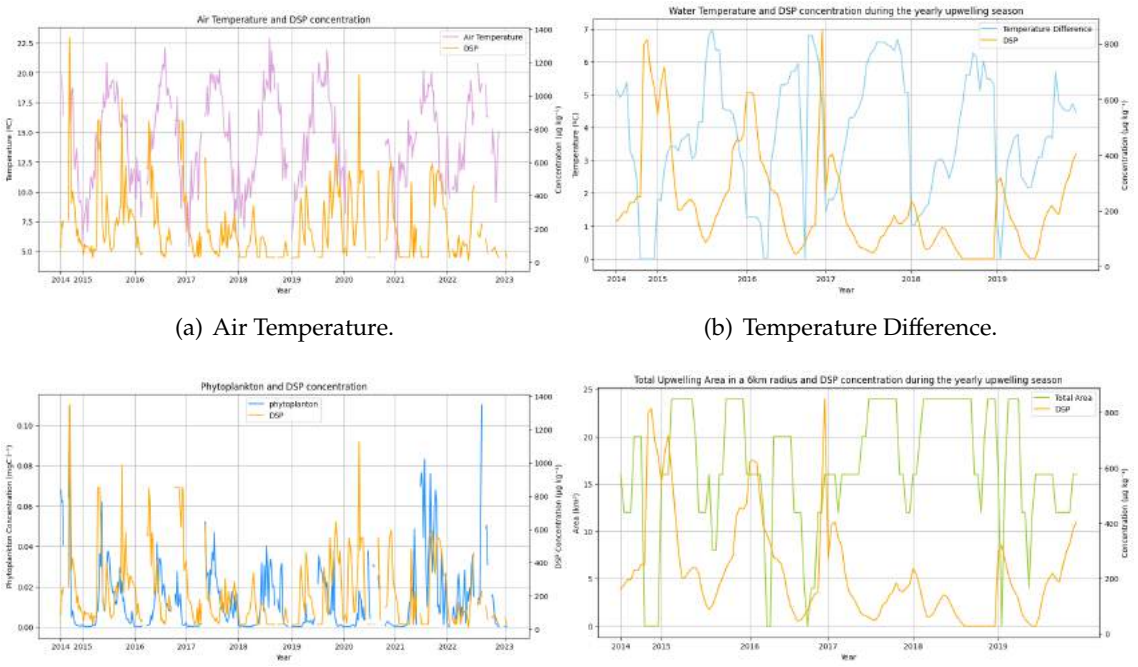


(a) Class distribution.

(b) Boxplot class balance.

Figure 4.12: L1 Carreço contamination class.

Following this analysis, it is important to study the remaining variables in order to identify yearly trends and understand any relation they might have with the DSP variable. In Figure 4.13 it is possible to observe the values for four variables along with DSP values. For every graph, the left y-axis denotes the focused variable while the opposite y-axis is related to the DSP variable, represented always in orange. Regarding the air temperature variable, displayed in Figure 4.13(a), it is possible to identify a yearly pattern with the exception of the year 2020 which is comprised of mostly missing data. For these four variables a clear relation between them and DSP is not present.



(a) Air Temperature.

(b) Temperature Difference.

(c) Phytoplankton.

(d) Total upwelling area.

Figure 4.13: L1 Carreço Environmental Variables.

For a further analysis on the relation between the DSP analysis and the other variables, the Pearson's correlations were calculated. This information is displayed as a heatmap, presented in Figure 4.14. For all variables the lagged values from 2 to 8 weeks, matching the window sizes studied in the models, were considered in order to study any possible long term effects on DSP variation. Due to the high number of variables and to achieve better readability, the y-axis was composed of only the DSP variables removing them from the x-axis. From the heatmap it is not possible to identify major correlations, either positive or negative, as the scale contains only low correlation values between -0.2 and 0.3. This can indicate a very complex dynamic on the L1 Carreço region where neither of the considered variables can perfectly explain the contamination behaviour.

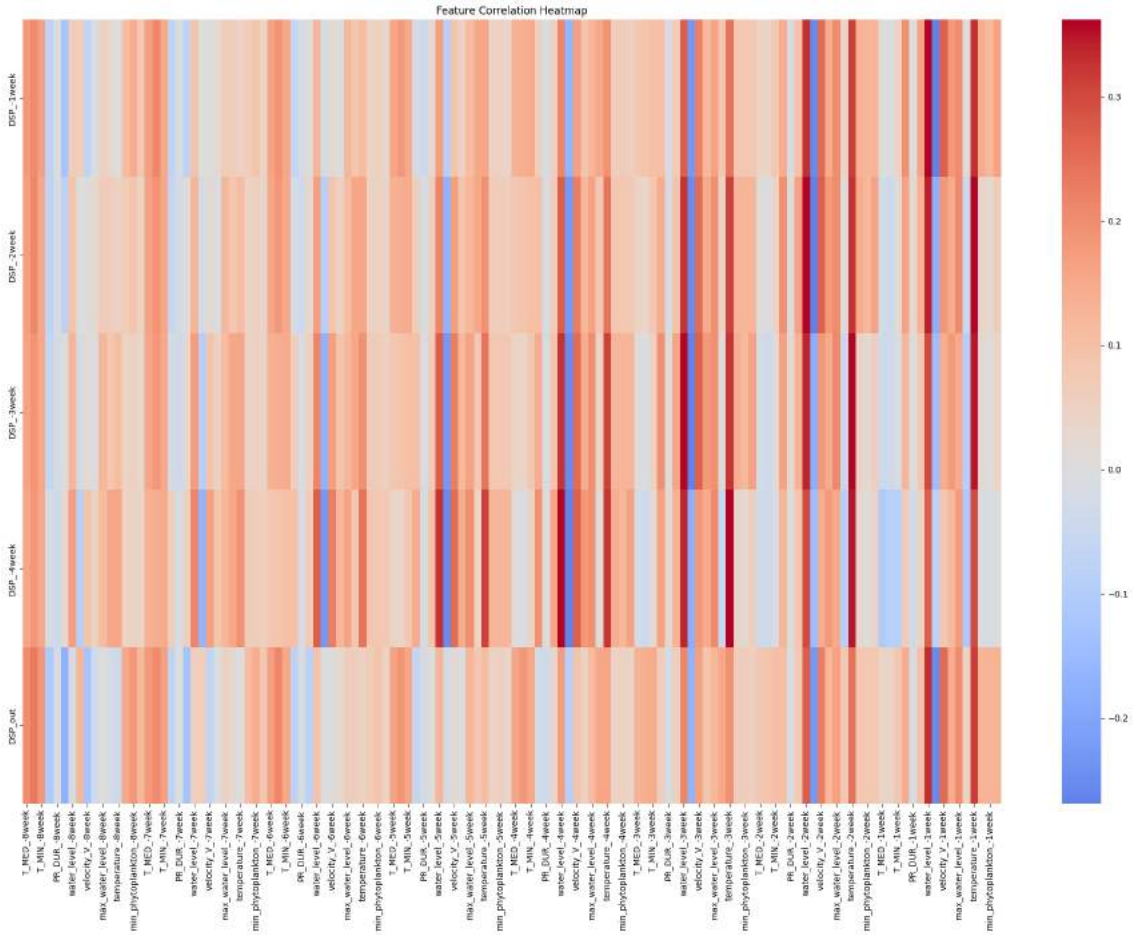


Figure 4.14: L1 Carreço Feature Correlation heatmap.

4.4 Random Forest / Support Vector Machine Models' Construction

In this section, the models' construction is presented, highlighting the main steps, followed by an extensive analysis of the hyperparameter tuning phase of these models.

4.4.1 Main Steps

In this subsection, the process of constructing the models is described in detail. As the construction of all models followed a similar process, the general procedure will be described, noting any differences when required. Constructing the models requires, firstly, a conversion of the data from a time-series format to a sliding window format, followed by a train-test split and data normalization phase, a walking forward cross-validation with hyperparameter tuning, sliding window size optimization and, finally, model testing and evaluation.

4.4.1.1 Sliding Window

The first step of the model construction process is to convert the data into a sliding window format. In this format, the weekly data of every variable is converted into a window containing data from a fixed number of sequential weeks, $\text{Week}_{\text{size}}$. For every group of valid $\text{Week}_{\text{size}} + 1$ samples, the first $\text{Week}_{\text{size}}$ samples become a window and the remaining one becomes the target variable. If any of the samples in $\text{Week}_{\text{size}} + 1$ is invalid, the window moves forward by one. The values of each variable within a window, become separate variables, representing the lagged instances. To determine the optimal $\text{Week}_{\text{size}}$ values between 2 and 8 were studied for all the models except the ones using datasets containing upwelling variables. In this case, only values between 2 and 5 were considered for the non upwelling variables, due to a lower number of samples.

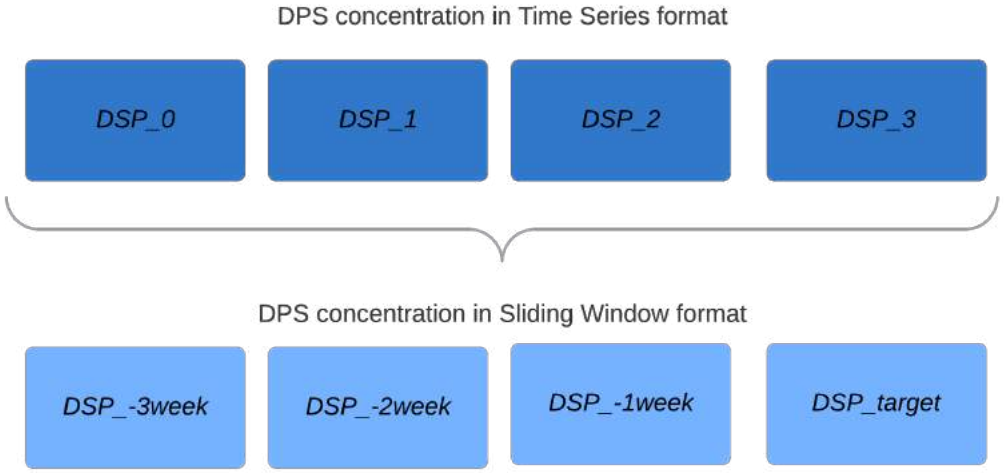


Figure 4.15: Sliding Window Conversion example for window size 3.

4.4.1.2 Train-Test Split and Data Normalization

Once the data is in sliding window format, it was split into train and test sets according to the 80:20 proportion. Due to the temporal nature of the data, the test set corresponds to the latest available data. Following the split, the variables were normalized using a standard Min-Max scaling formula that can be seen below:

$$X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}.$$

To keep the direct relation between the new lagged variables in every window, these were normalized in group, as opposed to individually, according to their original variable. The normalization scale was fitted to the training data and used to transform the training and testing data.

4.4.1.3 Walking Forward Cross-Validation

From the normalized training data, the hyperparameters, presented in Table 4.4 for RF models and in Table 4.5 for SVM and SVR models, were tuned with a walking forward approach using **sklearn's GridSearchCV and TimeSeriesSplit**. To achieve the best possible results, this technique was implemented using 5 folds and several scoring metrics. These metrics included, for classification models, balanced accuracy, average precision, recall and F1 Macro, with F1 Macro serving as the refit metric. For the regression models, RMSE, MSE, R^2 and MAE were used, with the latter being used for refit. These metrics were chosen based on the characteristics of the data, algorithms and approaches, focusing on class imbalance. Taking into consideration several metrics in this stage, previously described for each approach, allowed for a more detailed view of the results during the training, validation and testing phases.

Table 4.4: RF Hyperparameter Values

Hyperparameter	Values
n_estimators	100, 200, 300, 400
max_depth	5, 10, 12, 15, 20
min_samples_split	4, 5, 6, 7, 8
min_samples_leaf	4, 6, 8, 10, 12

Table 4.5: SVM and SVR Hyperparameter Values

Hyperparameter	Values	Notes
C	0.1, 0.5, 1, 2, 3, 5, 7, 10	-
gamma	0.1, 0.5, 1, 2, 3, 5, 7, 10	rbf and poly kernels
kernel	linear, rbf, poly	-
epsilon	0.01, 0.05, 0.1, 0.5, 1	SVR only
degree	2, 3, 4, 5	poly kernel only

4.4.1.4 Sliding Window Optimization

The previous tuning step was performed for every window size being considered, as described in Section 4.4.1.1. The results obtained from this step were used to choose the best window size for every model. Comparing the best obtained result when using each window size made it possible to select the size that achieved the best performance. For the execution of this step, two approaches were considered regarding the selection of the best window size. The first one, compared the best achieved results obtained for every window size while the second compared the average obtained result. Deciding on the first approach was justified as the results being considered were tied to the hyperparameter tuning. This means that choosing a window size that had the best average value but

not the best overall value would be using hyperparameters that did not achieve the best possible result.

4.4.1.5 Model Testing and Evaluation

Finally, from the tuned hyperparameters and the selected window size, the model was trained using the whole train set and tested with test set, described in Section 4.4.1.2. From the obtained results, several graphs were generated to facilitate the analysis and comparison between models. The produced graphs differed according to the algorithms and approaches used. RF models included information about feature importance extracted from the algorithm built-in mechanism. For classification models, the ROC and Precision-Recall curves, confusion matrices and misclassifications compared to the real DSP values were generated. For the regression models, the real and predicted DSP values were compared and confusion matrices were generated from converting the obtained data into contamination classes. For all the models, the values for the metrics being used, described in Section 4.4.1.3, were collected from the train, validation and test stages.

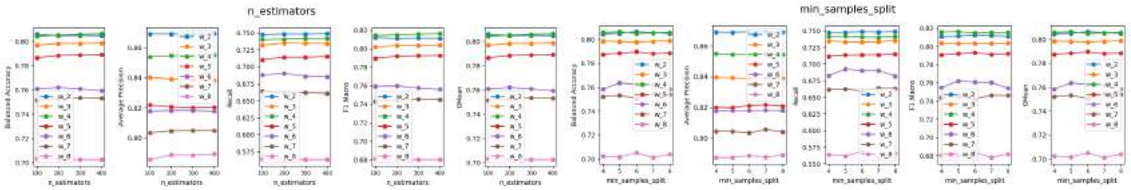
4.4.2 Hyperparameter Tuning

In this subsection, the results for the hyperparameter tuning phase, described in the previous section, are presented and compared between the developed models.

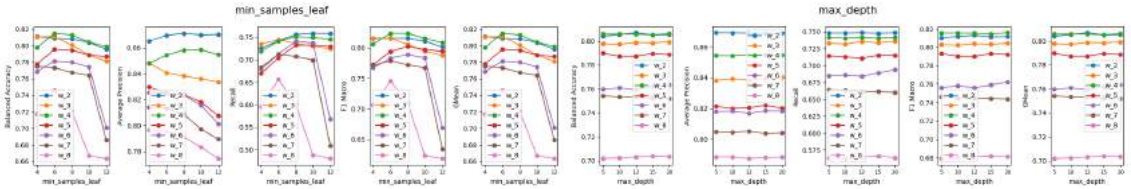
4.4.2.1 Classification : Random Forest

From the tuning phase of hyperparameters for the RF Classification models it is possible to obtain the average score in all considered metrics for all hyperparameter values being considered. In Figure 4.16, the average values for the L1-D model are presented. From the graphs, it is possible to observe that the smaller window sizes obtain higher average results, while the biggest window size obtained considerably lower average scores. Regarding the different tested values, with the exception of the max_depth parameter where a clear decrease in average performance happens for the bigger values in the tested range, the average scores remain very similar across the whole range. The averaging of the scores and the display of different window sizes does not allow an easy identification of subtle variations in performance when using the specified values nor conclusions to be drawn about the best window size.

Following the hyperparameter tuning, the optimal window size for each model was chosen, through the process described in Subsection 4.4.1.4. The best results for each window size for the L1 Carreço region are displayed in Figure 4.17. From these results, it is possible to observe that the L1-D model obtains better results in all window sizes than the other models, which all presented very similar results. It is also possible to identify that all models obtained better results for the smaller window sizes, with all models except the L1-D presenting a sharp decline in performance with the increase in window size.



(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure 4.16: RF hyperparameters tuning for the L1-D model.

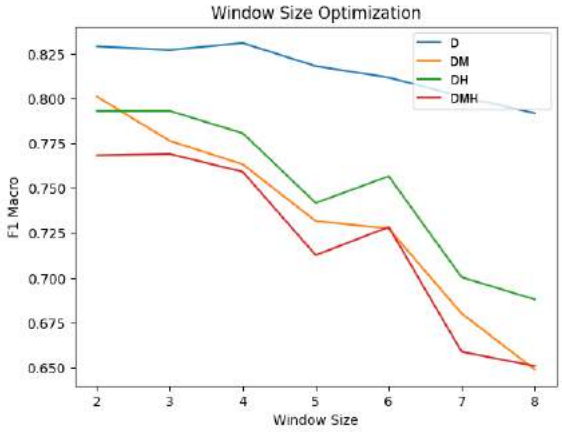


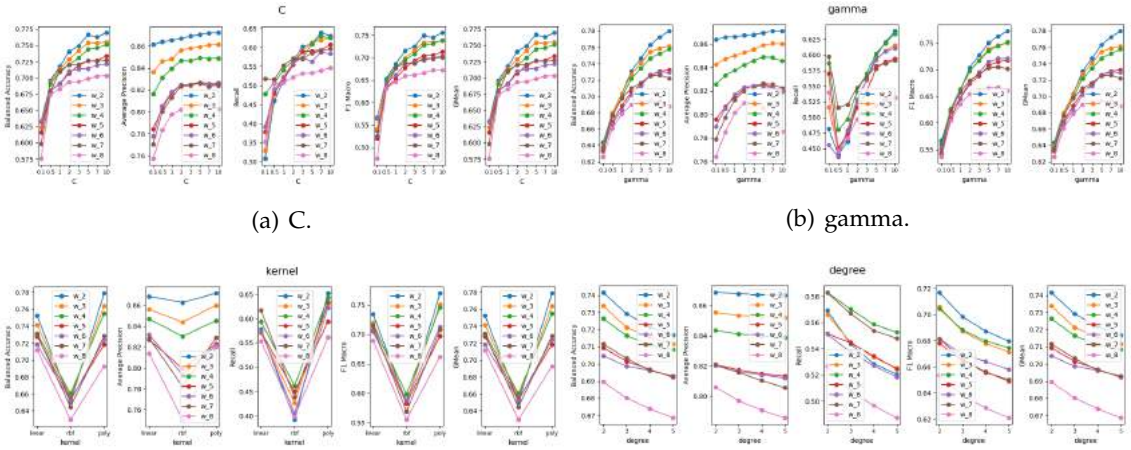
Figure 4.17: L1 Carreço Window Optimization for RF Classification models.

The results obtained on the other zones by the RF Classification models at this phase mostly follow the previous observations. Nonetheless, it is possible to identify some differences in the tuning phase regarding the average performance of the middle window sizes that vary between being similar to results for the smaller or bigger window sizes. It is also possible to observe, for a few models as L5b-DM and RIAV1-D, bigger window sizes as 5 and 8, respectively, obtaining the best results. For the upwelling models in the different regions it is possible to observe a bigger difference between average results of different window sizes, however, for the window optimization, the best results were more constant across window sizes.

4.4.2.2 Classification : Support Vector Machine

For the SVM classification models, during the hyperparameter tuning phase, exemplified in Figure 4.18 with average performance values for the L1-D, a clear division between

window size is not as evident as previously observed in the RF classification models. For these models, the tested window sizes obtained similar average scores for many hyperparameters and metrics. It is also important to notice, that for these models, there is an increased variation in average scores between values of a hyperparameter. In Figure 4.18(c), although the rbf kernel averages scores quite lower than the other tested kernels it is still used for many of the best results obtained across window sizes. This evidences a high variation in results when using the rbf kernel, that can be explained by poorer combinations with the other hyperparameters for this kernel.



(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure 4.18: SVM Classification hyperparameters tuning for L1 Carreço DSP dataset.

In Figure 4.19, the best results in validation when using each window size for all tested models are presented for the L1 Carreço region. For this region all models obtained the best results using a 2-week window size, presenting a considerable decrease in performance for the other window sizes.

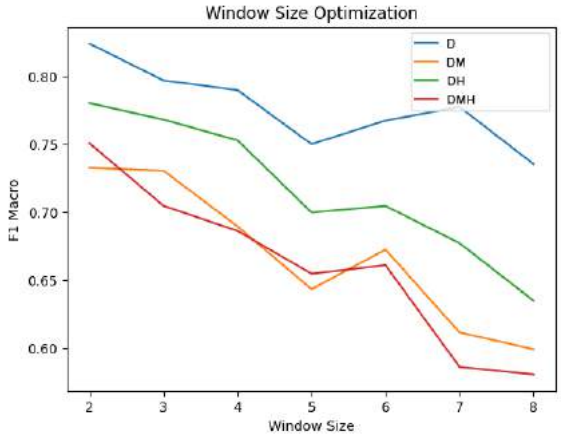
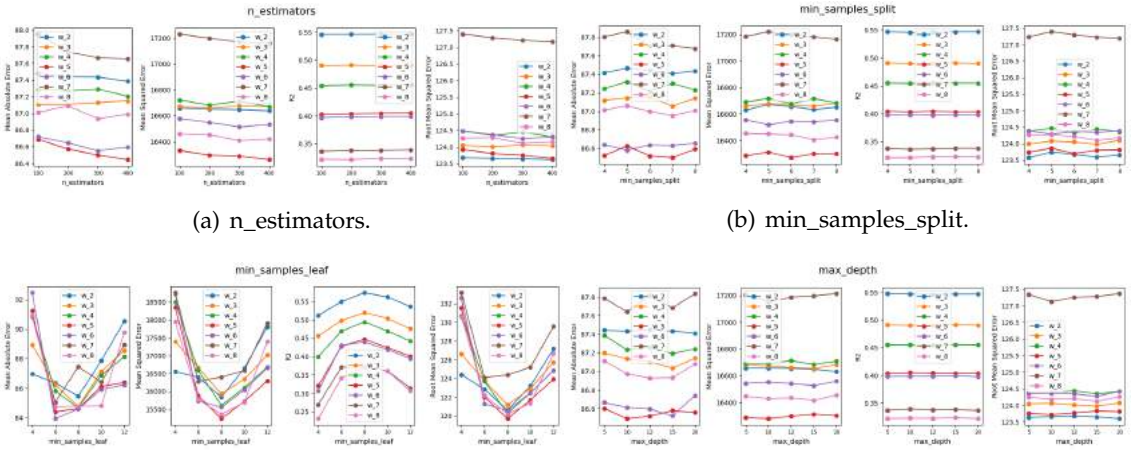


Figure 4.19: Window Optimization for L1 Carreço SVM Classification.

For the other regions, it is equally possible to identify high variations in average scores between values and small variations across window sizes. Contrary to the L1 Carreço region, the best results in the other zones are more similar across window sizes. The results for the upwelling models are similar to the previous observations, showing also a significant variation in average scores between different values. However, contrary to the non-upwelling models, the differences between window sizes in the average scores are clearer. Additionally, for the upwelling models, the best results between window sizes are more similar with slightly bigger window sizes obtaining the best results more frequently than in the previous models.

4.4.2.3 Regression : Random Forest

For the RF regression models, from the average scores during the hyperparameter tuning phase, exemplified in Figure 4.20 for the L1-D model, it is possible to observe that some of the biggest window sizes present some of the lowest average errors, with the exception of the R^2. This differs from the Classification models, where the smaller window sizes would mostly obtain slightly better average scores than their bigger counterparts. As previously observed for the RF Classification models, it is also clear that most scores remain constant between values, with the exception of the hyperparameter min_samples_leaf, where better average performances are clearly obtained for the values in the middle of the tested range.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure 4.20: RF Regression hyperparameters tuning for L1 Carreço DSP dataset.

Regarding the window optimization phase, represented in Figure 4.21, it is possible to observe that the models using HWP variables obtain very similar results. Nonetheless, the best performance is clearly obtained in this phase by the L1-D model. It possible to observe that for most models, the best results of each window size are very similar, leading to a wide range of window sizes being preferred by the different models.

For the other zones, the bigger window sizes present poorer average results and an

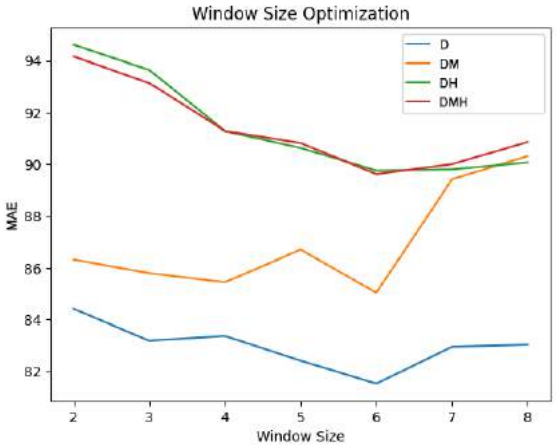


Figure 4.21: Window Optimization for RF Regression in L1 Carreço.

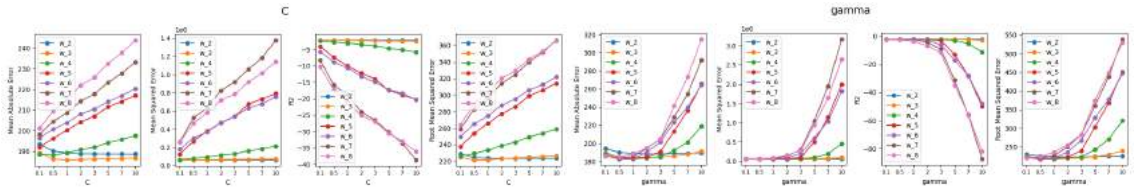
increase in their lowest errors, with the smallest window size always being considered optimal. The previous remarks were even more clearly evidenced for in all upwelling models, with the smallest window size always presenting the best results, except for the L1 Carreço upwelling models where all window sizes presented similar results.

4.4.2.4 Regression : Support Vector Regression

The hypertuning for the SVR Regression models present some similarities to the SVM Classification models. In Figure 4.22, the average results of the hypertuning phase are presented for the L1 Carreço region. Similarly to the SVM models, from these results, it is possible to observe a high variation in average scores between the range of values, while presenting considerably low variations between window sizes. Additionally, it is possible to identify that, contrary to the other regression models, the bigger window sizes present some of the lowest average error scores.

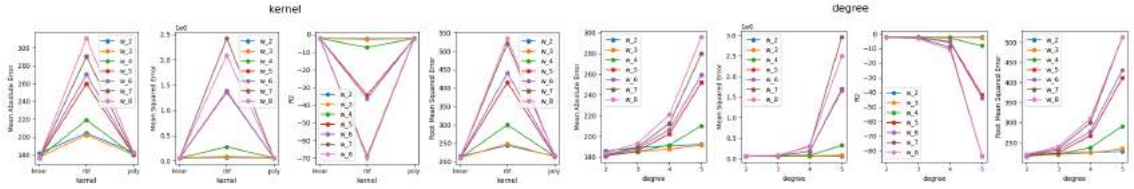
Considering the best scores obtained using the different window sizes by the models, present in Figure 4.23 for the L1 Carreço region, it is possible to observe that the lowest error scores were mostly obtained by the smaller window sizes, often presenting a considerable increase for the bigger window sizes.

The results obtained for the other zones are in accordance with the previous observations, not registering any major differences. The same conclusions were achieved regarding the upwelling models in all regions.



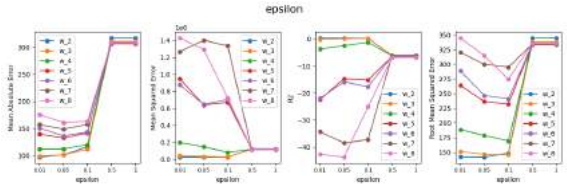
(a) C.

(b) gamma.



(c) kernel.

(d) degree.



(e) epsilon.

Figure 4.22: SVR hyperparameters tuning for L1 Carreço DSP dataset.

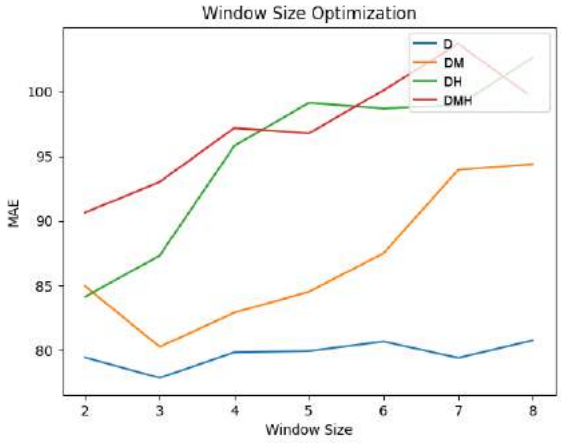


Figure 4.23: Window Optimization for SVR Regression.

5

Results discussion and analysis

In this chapter the results obtained from the developed classification and regression models will be discussed and compared. These models were developed using Python and the pandas, numpy, seaborn, and scikit-learn libraries. The models' source code, as well as all the results and plots generated, are available in a public repository[56]. Firstly, in Section 5.1, the results for the classification approach will be presented and discussed. Secondly, the regression results will be analysed following the same structure. Finally, a comparison between approaches is presented in Section 5.3. To be concise, the discussion will be focused on the results of the L1 Carreço region with significant differences in the results of other regions being noted.

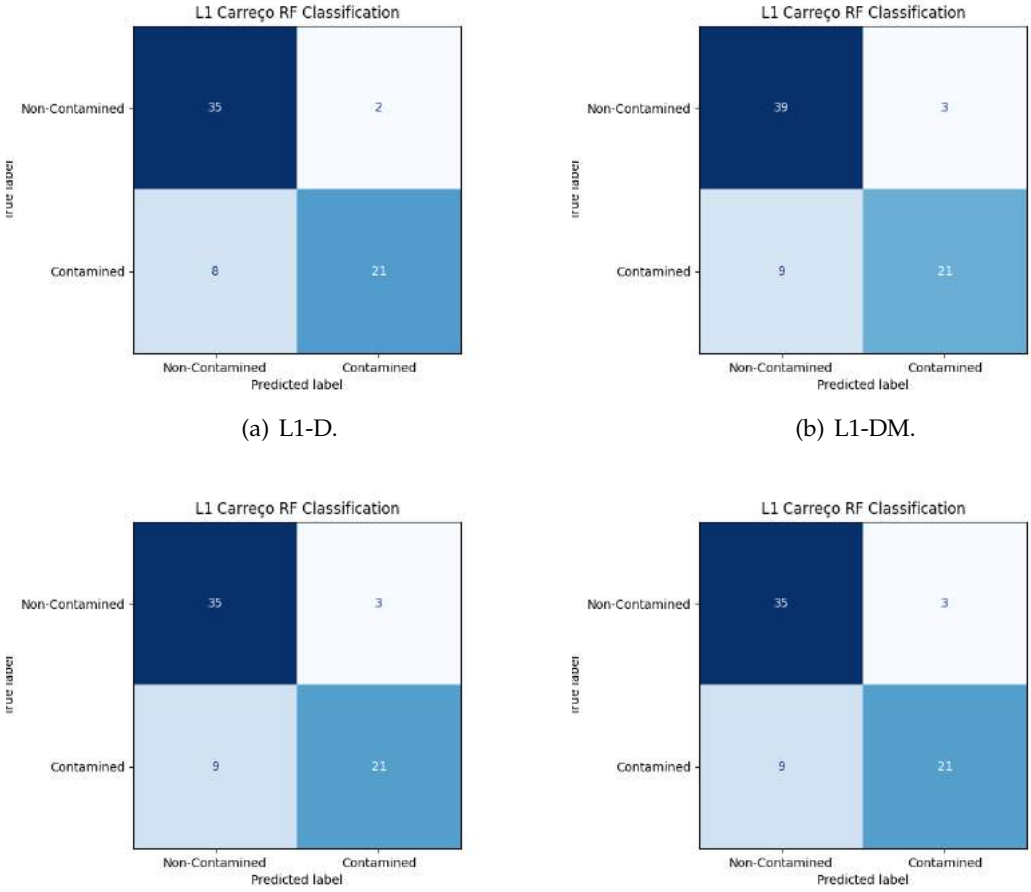
5.1 Classification

The classification results will, firstly, be analysed for the models using the RF algorithm for each dataset and their performance compared. Following this analysis, the SVM results will be discussed in the same manner. Finally, a comparison between the algorithms being used in this approach is presented.

5.1.1 Random Forest

The obtained results will be displayed as confusion matrices in order to easily analyse and compare correct and incorrect classifications between the developed RF Classification models. From the results obtained by these models for the L1 Carreço region, displayed in Figure 5.1, it is possible to observe that the models performed similarly, as all of them achieved the same quantity of True Positive (TP) and True Negative (TN), with the exception of the L1-DM model, that showed a discrepancy in TN quantity. To further analyse the differences between the models, in an early stage, only the TPR and TNR were used to compare their performances.

- For the L1-D model, it is possible to identify that, besides a similar TP and TN to other models, it shows the lowest amount of False Positive (FP) and False Negative



(a) L1-D.

(b) L1-DM.

(c) L1-DH.

(d) L1-DMH.

Figure 5.1: RF Classification Confusion Matrices for L1 Carreço.

(FN). The lower amount of misclassifications leads to a TPR of 72.41% and a TNR of 94.59%.

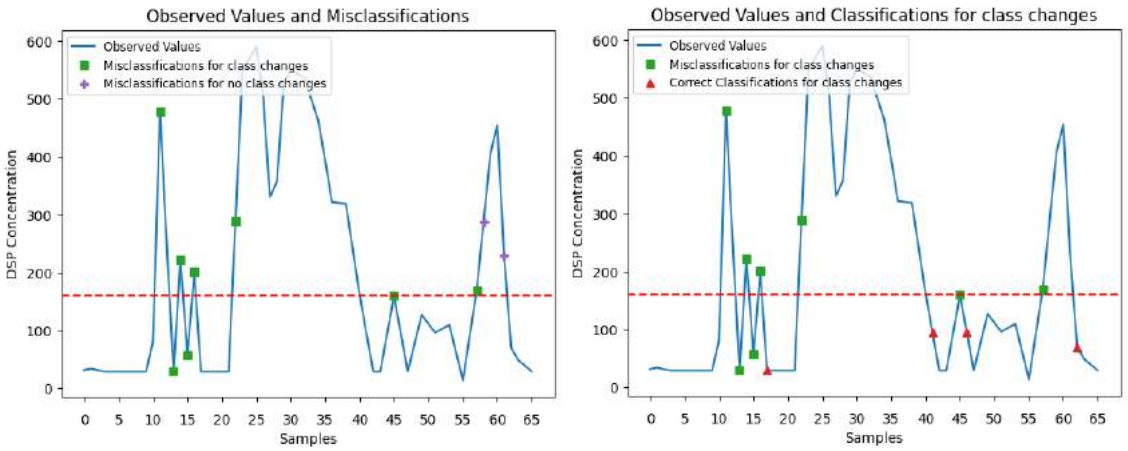
- The L1-DM model presents an increase in FN when compared with the L1-D model, lowering its TPR to 70.0%. Due to selecting a lower window size, this model contains more samples than the other models, presenting a much higher quantity of negative samples from which a TNR of 92.86% is obtained.

- The L1-DH and L1-DMH present the same prediction results, having the same TP and TN as the L1-D model but registering an increase in FN and FP. This change results in a decrease in TPR, to 70.0%, matching the performance on positive samples of the L1-DM model and a decrease in TNR, to 92.11%.

From the previous results, it is possible to observe that the L1-D model obtains the highest TNR and TPR of all the models in the comparison. Additionally, besides analysing the results in the confusion matrices, it is helpful to study the scenario that lead to the

misclassifications. This was performed by examining the classification results and comparing them to the real DSP values. In Figure 5.2, it is possible to observe the real DSP values along with marked classifications obtained for the L1-D model. The marked classifications in Figure 5.2(a), represent all the misclassifications and are split into 2 categories; the misclassifications for class changes, that express misclassifications on a sample that presents a different contamination class to its predecessor; the misclassifications for no class changes that relate to misclassifications on samples that indicate the same contamination class of it predecessor. Similarly, the marked classification in Figure 5.2(b) depict all the misclassifications for class changes while also showing the correct classifications when a class change occurs.

From the displayed graphs, it is possible to identify that 80.0% of the misclassifications of the L1-D model happen on samples with contamination class change. As is it possible to observe, these misclassifications, marked as green squares always occur on sudden changes in DSP contamination with many of the misclassified samples containing concentration values close to the contamination threshold. In the graph in Figure 5.2(b), 66.67% of the class changes are shown to result in misclassifications. From the analysis of both graphs, it is possible to conclude that DSP rapid variations that result in sudden changes of the contamination classes are critical for the model and are its main struggle.



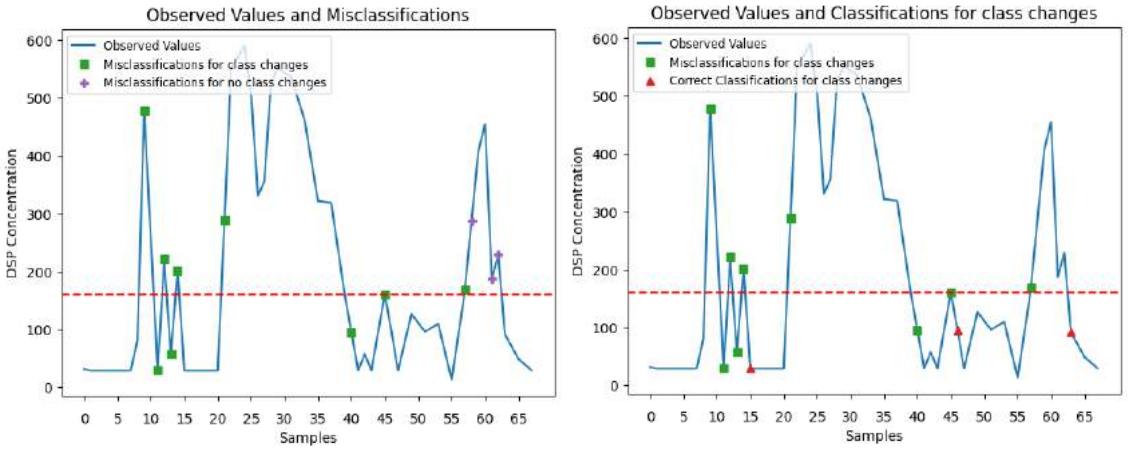
(a) Misclassifications.

(b) Class changing samples.

Figure 5.2: L1-D classifications for class changing samples.

From studying the results of class changing samples it is possible to better understand the slight decrease in performance of the other models when compared with the L1-D model. Analysing the classification results on class changing samples for the L1-DH model, presented in Figure 5.3, it is possible to observe that its additional misclassifications, that lead to a lower TPR and TNR, result from a misclassification for a class changing sample that was correctly classified by the L1-D model and an additional misclassification on a non-class changing sample resulting from the increased complexity due to a higher

number of samples.

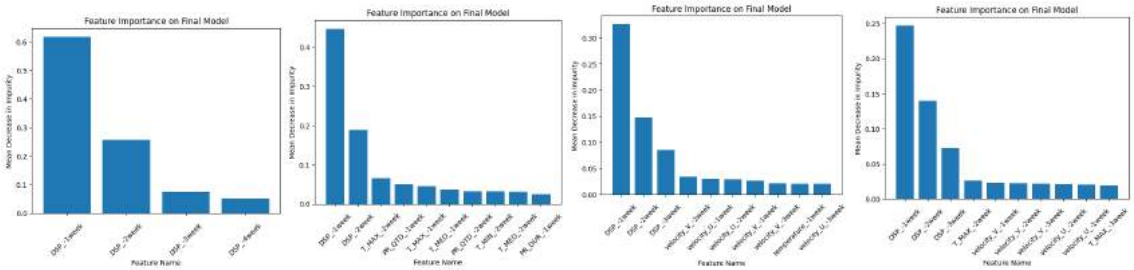


(a) Misclassifications.

(b) Class changing samples..

Figure 5.3: L1-DH classifications for class changing samples.

For these models, the feature importance of the used variables is available through the built-in feature relevance of the RF algorithm, results which are present in Figure 5.4. From these results, it is possible to observe that across all models, the variable representing the DSP of the previous week is considered much more relevant than the other features, with non-DSP variables presenting very low relevance values. The slightly poorer results for the models using more than the DSP variable, can be justified by the decrease in importance of the DSP of the previous week due to the increase in number of variables used, that, by itself present very low importances but that together worsen the model and the overall importance of the main variable.



(a) L1-D.

(b) L1-DM.

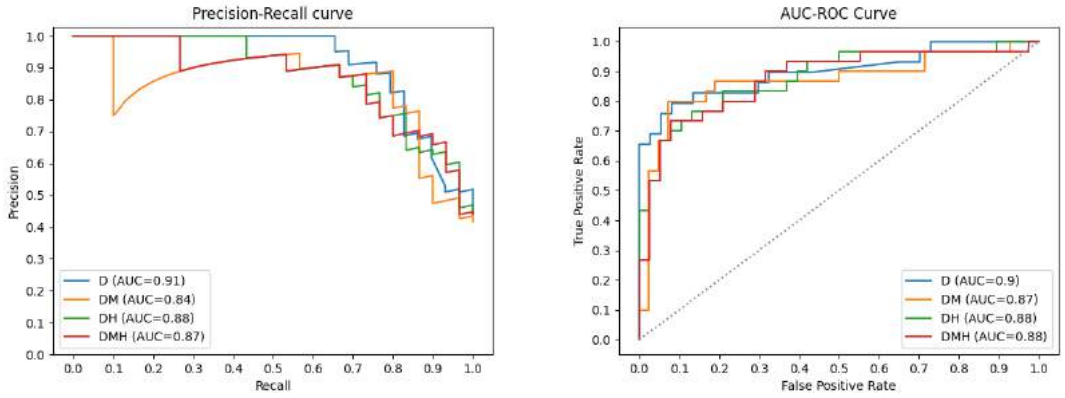
(c) L1-DH.

(d) L1-DMH.

Figure 5.4: L1 Carreço RF Classification feature importance.

In Figure 5.5, it is possible to observe the Precision-Recall and AUC-ROC curves for the L1 Carreço region. The probabilities obtained for every model and represented in the curves for the different thresholds confirm that the models perform very similarly, with the preferable model varying across thresholds. Nevertheless, in both graphs the L1-D

model obtains a slightly higher AUC, meaning that its deemed as the most capable model in terms of classification.



(a) PR-REC. (b) ROC-AUC.

Figure 5.5: L1 Carreço RF Classification models curves.

In Table 5.1, the train, validation and test results for all the selected metrics are presented. From the results, it is possible to identify that the model using the DSP dataset presents better validation and test results across all metrics when compared to the other models, despite having the worst scores for the train set. These scores are in accordance with the previous analysis and supports the claim that the L1-D model has a slightly better classification capability than the other models, indicating that the model using the simplest data can achieve better results.

Table 5.1: RF Classification metrics for L1 Carreço

Set	Metric	Dataset			
		D	DM	DH	DMH
Train	Balanced Accuracy	0.8704	0.9366	0.9121	**0.9510**
	Average Precision	0.9601	**0.9908**	0.9753	0.9903
	F1 Macro	0.8748	0.9371	0.9131	**0.9529**
	Recall	0.8103	**0.9349**	0.8812	0.9347
Validation	Balanced Accuracy	**0.8138**	0.7710	0.7282	0.7706
	Average Precision	**0.8543**	0.8077	0.8001	0.7920
	F1 Macro	**0.8174**	0.7684	0.7241	0.7615
	Recall	**0.7318**	0.6925	0.6607	0.7201
Test	Balanced Accuracy	**0.8350**	0.8143	0.8105	0.8105
	Average Precision	**0.9117**	0.8449	0.8824	0.8693
	F1 Macro	**0.8413**	0.8222	0.8157	0.8157
	Recall	**0.7241**	0.7000	0.7000	0.7000

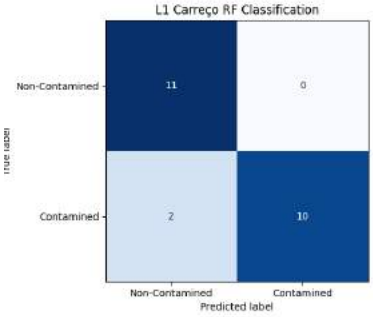
The results obtained by the RF Classification models for the other test zones are similar to the results discussed for the L1 Carreço, achieving, however, a slightly better performance, specially regarding their power to correctly classify the positive class. It

is also important to notice that, contrary to the preference of the simplest model in the previous analysis, the inclusion of additional variables seem to enhance performance in the other regions.

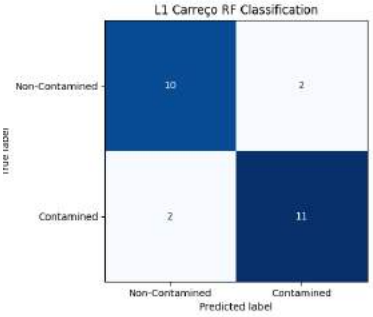
- For the L2 Leça da Palmeira zone, both models, L2-D and L2-DM, achieve the same performance, registering a TPR of 81.82% and a TNR of 92.11%. Analysing the PR and ROC curves, it is possible to observe very similar probabilities between the two models, with the L2-DM model obtaining a slightly higher AUC in both curves.
- From the predicted classifications for the L5b Caparica region, it is possible to observe that the model using meteorological variables achieves a better performance, with a TNR of 93.75%, slightly higher than the 91.67% obtained by the L5b-D model. Regarding positive samples, the L5b-DM model also ranked higher with a TPR of 88.57% surpassing the 81.58% of the the L5b-D model. The better performance of the L5b-DM model is reflected in its higher F1 Macro score when compared to the L5b-D model, as presented in Table C.2. Uniquely, this model showed a higher feature importance for the DSP concentration of two weeks before rather than the usual DSP of the previous week.
- For the RIAV1 Triângulo region, the RIAV1-DH and RIAV-DMH models present the same TPR at 81.58% with the latter having a TNR of 75.68%, surpassed by the 78.38% of the former. Slightly poorer TPR and TNR were obtained by the RIAV1-DM model, at 76.32% and 72.97%, respectively. Using a different window size, containing more positive samples and much less negative ones, the RIAV1-D model performed very well for the positive classes, presenting the highest TPR of all models at 86.05%, failing, however, at correctly classifying the negative samples with a TNR of 63.18%. In Table C.3, it is possible to confirm the overall better performance of the RIAV1-DH model, that presents a more balanced performance considering both classes, leading to the highest F1 Macro and balanced accuracy scores.

5.1.1.1 Upwelling Results

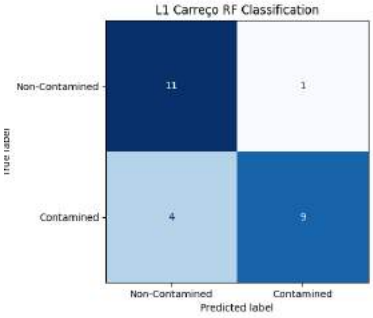
The models using upwelling-focused datasets were, firstly, compared separately due to their higher number of models and their different temporal resolution, making it hard to perform the previous analysis along with the non-upwelling models. The obtained results from these models are displayed in the confusion matrices present in Figure 5.6. In order to examine the impact of the upwelling variables, the confusion matrices will be analysed in pairs of upwelling and non-upwelling variables.



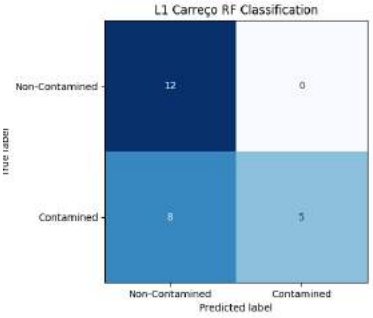
(a) L1-UP-D.



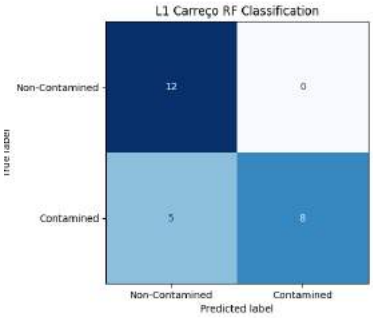
(b) L1-UP-DU.



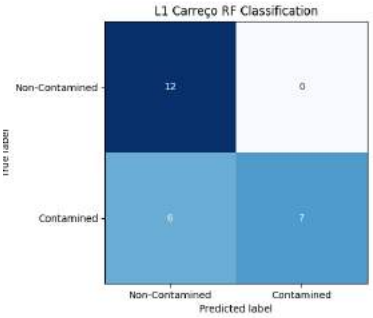
(c) L1-UP-DM.



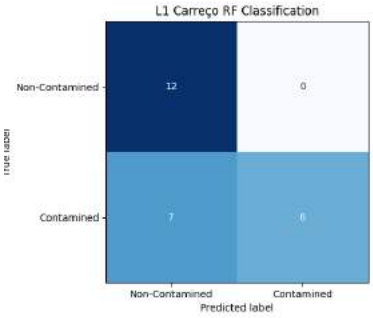
(d) L1-UP-DMU.



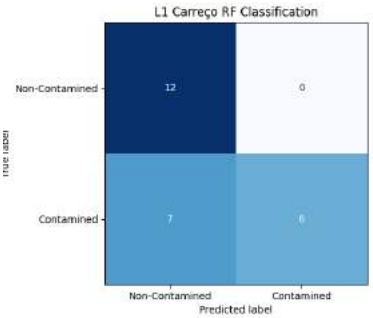
(e) L1-UP-DH.



(f) L1-UP-DHU.



(g) L1-UP-DMH.



(h) L1-UP-DMHU.

Figure 5.6: Confusion Matrices for RF Classification Upwelling models.

- Starting with the L1-UP-D and L1-UP-DU it is possible to observe that the model including upwelling variables presents a slight increase in TPR, at 84.62% when compared with L1-UP-D's TPR of 83.33%. This slight increase comes, however, with a considerable reduction in TNR to 83.33% from the 100.00% obtained by the L1-UP-D model as it correctly classified every non-contaminated sample. Nevertheless, these two models manage to achieve much better results than any of the other models.

- The models L1-UP-DM and L1-UP-DMU managed to achieve very high TNR values of 91.67% and 100.00%, respectively. However, both models achieved much poorer results on the positive class with a TPR of 69.23% for the L1-UP-DM model, with the model including upwelling variables being much worse at classifying the positive samples with a TPR of 38.46%.

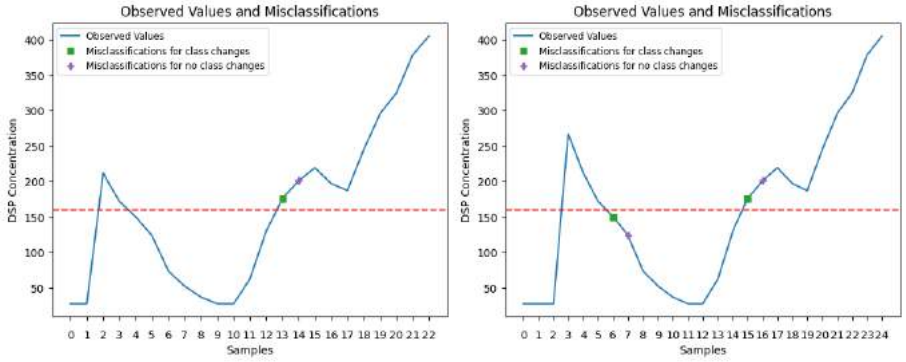
- For the L1-UP-DH and L1-UP-DHU models, a similar performance on the negative class as the previous models was achieved, with a TNR of 100.00% for both models. Similarly to the previous models, these also struggled with the positive class, presenting, respectively, a TPR of only 61.54% and 53.84%.

- Finally, the L1-UP-DMH and L1-DMHU models, despite achieving a perfect classification of the negative samples, with a TNR of 100%, failed at correctly classifying most of the positive samples, with a low TPR of 46.15%

From these results it is possible to observe that the simplest models obtained the best results. It was also possible to identify that the addition of both HWP and meteorological variables, resulted in a decrease in performance. Regarding the impact of the upwelling variables, further research is needed as, when comparing between the pairs, it was not possible to observe a clear impact on performance.

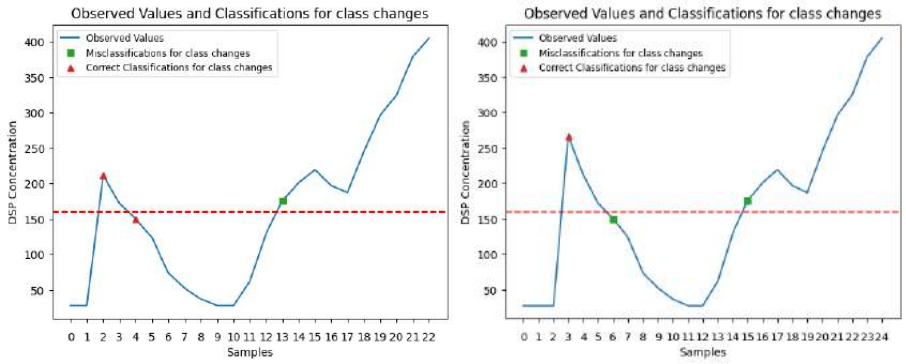
In order to further study the impact of these variables it is necessary to analyse the classifications on class changing samples between the pairs. In Figure 5.7, the misclassifications of the L1-UP-D and L1-UP-DU models are categorized based on the presence of a class change. From these graphs it is possible to observe that the first model obtains a misclassification on a class change sample followed by another misclassification on the following sample, indicating that the model failed to immediately adjust to the first misclassification. For the latter model, the same behaviour is shown with an additional occurrence of this double misclassification phenomenon.

Regarding all class changing samples, present in Figure 5.8, it is possible to observe that both models manage to correctly classify the first class changing sample, with the L1-UP-D model also managing to correctly identify the second class change, which is misclassified by the L1-UP-DU. These differences indicate a better performance by the L1-UP-D model making it the most desirable of the two.



(a) Missing distribution. (b) Missing periods.

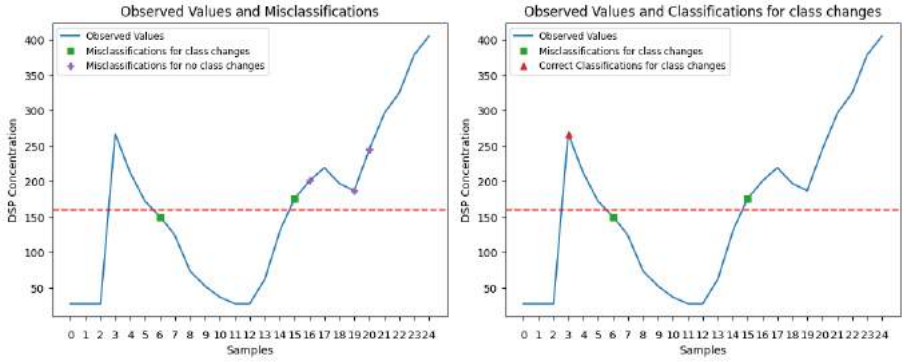
Figure 5.7: Misclassifications for for RF L1-UP-D and L1-UP-DU.



(a) Missing distribution. (b) Missing periods.

Figure 5.8: Class changes for RF L1-UP-D and L1-UP-DU.

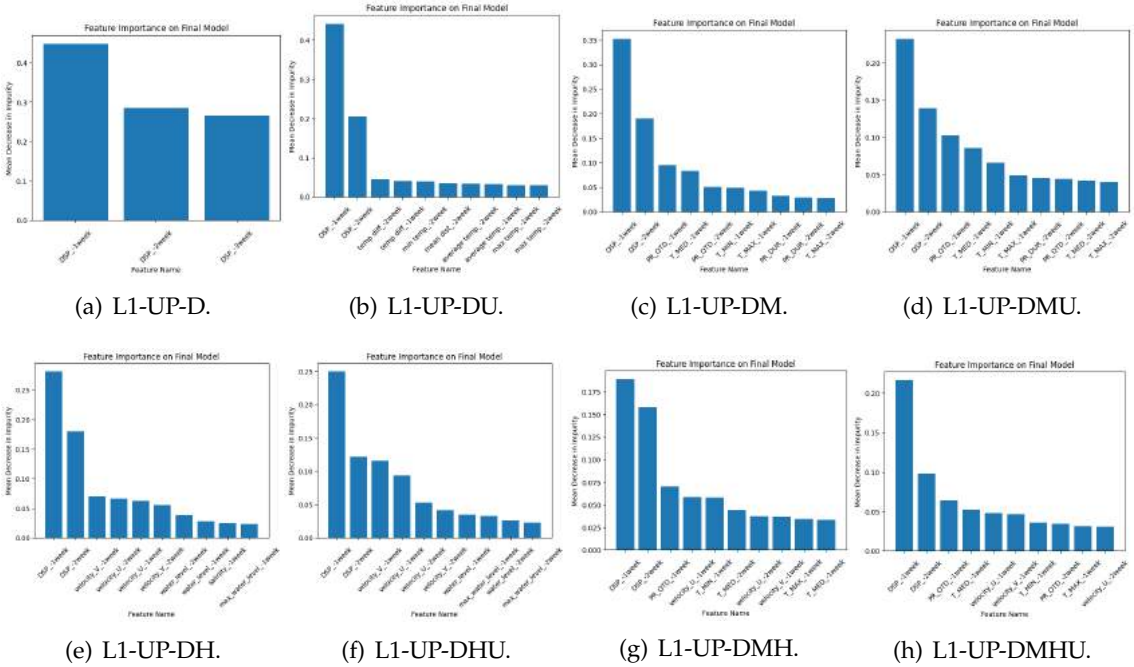
To better understand the decrease in performance of the other upwelling models it is necessary to also analyse their misclassifications and compare them with the previously discussed ones. In Figure 5.9(a), the class changing samples and the misclassification for the L1-UP-DM model are presented. From these graphs, it is possible to identify that this model obtains the same classification results on class changing samples as the L1-UP-DU model. The misclassification pattern observed in the previous models, where a misclassification on a class change would be always followed by another misclassification, is not replicated for the first misclassification of this model, being present in the second one. Additionally, two extra misclassifications are obtained for non-class changing samples making this model less desirable and justifying its lower performance.



(a) Missing distribution. (b) Missing periods.

Figure 5.9: RF L1-UP-DM class changes and misclassifications.

From the RF algorithm being used in these models, the feature importance graphs were extracted and are presented in Figure 5.10. From these, it is possible to observe that, similarly to the non-upwelling models previously analysed, the DSP concentration of the previous week is always the most relevant feature. It is also possible to observe that the models that achieved a better performance had an increased relevance for this variable. Comparing the importance of this variable in each model, a better performance can be related to the increase in relevance of this variable.



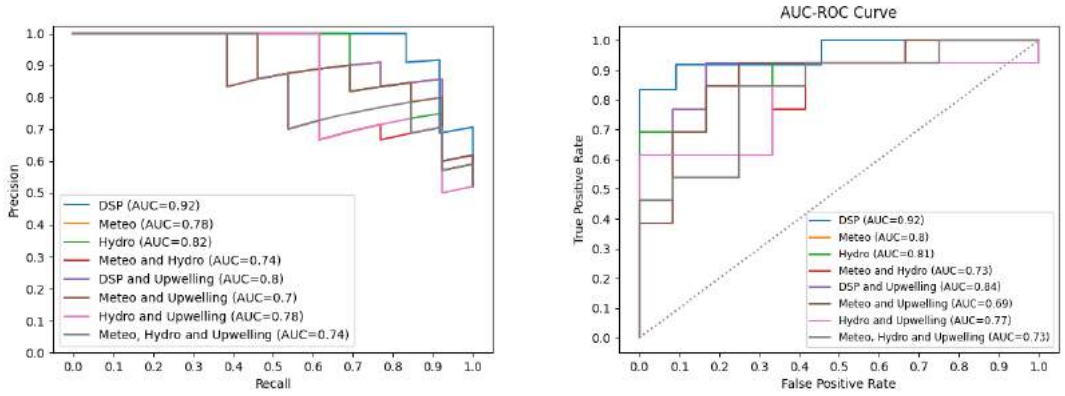
(a) L1-UP-D. (b) L1-UP-DU. (c) L1-UP-DM. (d) L1-UP-DMU.

(e) L1-UP-DH. (f) L1-UP-DHU. (g) L1-UP-DMH. (h) L1-UP-DMHU.

Figure 5.10: L1 Carreço RF Classification Upwelling Feature Importnace.

Analysing the curves in Figure 5.11, the preference for the L1-UP-D model is reinforced, ranking higher in both curves, with the highest AUC. From the obtained scoring metrics,

present in Table 5.2, it is possible to confirm that the L1-D is superior to the other models, since it presents the highest F1 Macro, balanced accuracy and average precision scores for the test set. As analysed before, the L1-UP-DU presents the highest recall having the best performance on positive samples.



(a) PR-Curve. (b) AUC-ROC Curve.

Figure 5.11: L1 Carreço RF Classification Upwelling curves.

Table 5.2: RF Classification Upwelling metrics for L1 Carreço

Set	Metric	Dataset							
		D	DU	DM	DMU	DH	DHU	DMH	DMHU
Train	Balanced Accuracy	0.8778	0.8625	0.9281	0.9753	0.9602	0.9420	0.9905	**0.9943**
	Average Precision	0.9936	0.9979	0.9973	0.9992	0.9995	0.9992	**0.9997**	**0.9997**
	F1 Macro	0.8948	0.8892	0.9388	0.9804	0.9695	0.9578	0.9905	**0.9952**
	Recall	0.9726	0.9837	0.9761	0.9924	0.9924	0.9924	0.9924	**1.000**
Validation	Balanced Accuracy	**0.8752**	0.7624	0.8039	0.8198	0.6858	0.6559	0.7880	0.7782
	Average Precision	0.8029	0.8177	0.8805	0.8726	0.8931	0.8206	**0.9096**	0.8850
	F1 Macro	**0.8532**	0.6614	0.7227	0.7312	0.5625	0.5097	0.7005	0.6871
	Recall	0.8889	**1.000**	0.9524	0.9524	**1.000**	**1.000**	0.9524	0.9524
Test	Balanced Accuracy	**0.9167**	0.8397	0.8045	0.6923	0.8077	0.7692	0.7308	0.7308
	Average Precision	**0.9203**	0.7960	0.7831	0.7046	0.8154	0.7785	0.7415	0.7415
	F1 Macro	**0.9129**	0.8397	0.7987	0.6528	0.7947	0.7500	0.7029	0.7029
	Recall	0.8333	**0.8462**	0.6923	0.3846	0.6154	0.5385	0.4615	0.4615

Analysing the results for the other test zones, it is possible to observe that the L5b Caparica and L7c2 Porto de Mós regions obtain similar results to the L1 Carreço region, while better results are obtained in the L2 Leça da Palmeira for all the developed models.

- For the L5b Caparica region, the samples in the test set are composed of mostly positive samples containing a very reduced amount of negative samples. Both L5b-UP-D and L5b-UP-DU struggle with these negative samples presenting a TNR of 50.00% and 0.00%. These results, however, can not be considered representative of

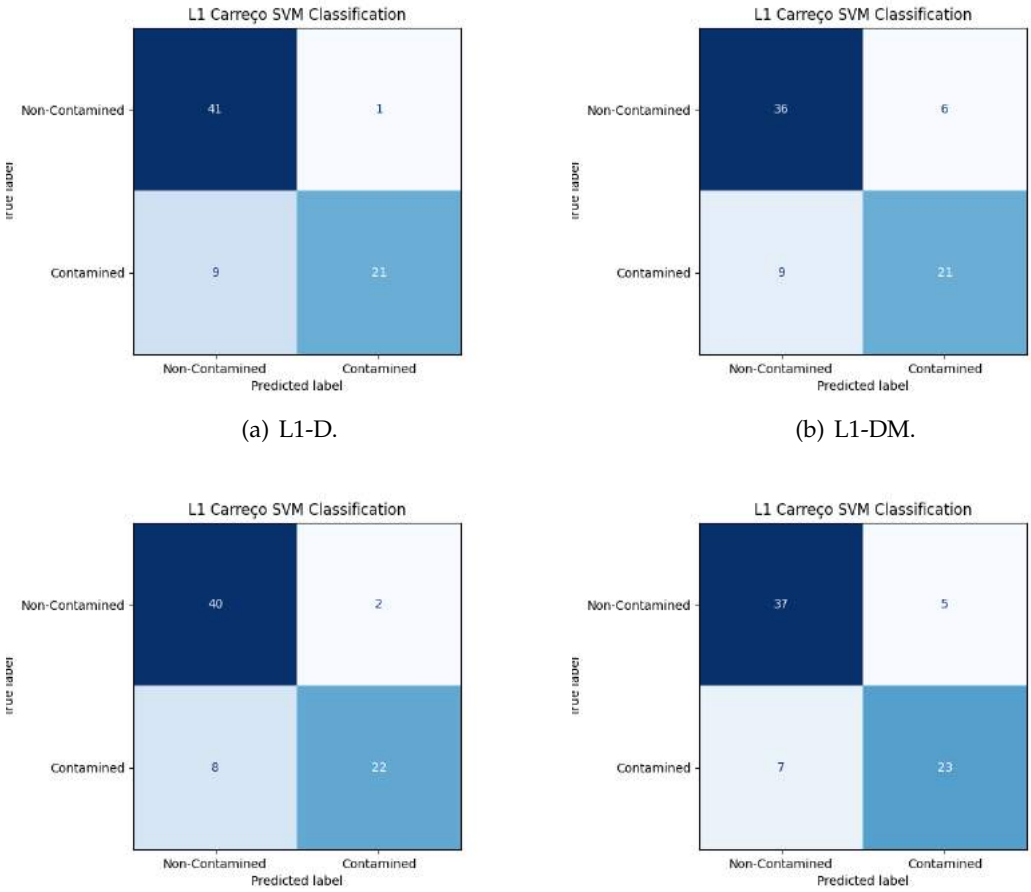
the classification power of these models for negative samples as only two negative samples were classified. Regarding, the positive classes, both these models present a good classification capability, with TPR of 88.00% and 84.00%. Similarly to the results obtained in the L1 Carreço region, the inclusion of meteorological variables came with a heavy decrease in performance, with L5b-UP-DM and L5b-UP-DMU presenting a much poorer classifying ability for positive samples, with the latter completely failing to correctly classify any of the positive samples. These conclusions strengthen the claims made for the L1 Carreço region, also showing a good prediction capability for the simplest model, with a small decrease in performance when including upwelling variables. In both zones, the addition of meteorological variables correlated to a decrease in performance. This decrease in performance is clearly evidenced in the scoring metrics in Table C.5.

- From the results obtained in the L7c2 Porto de Mós region, similar patterns are observed, although not as pronounced. All the models obtained the same TPR for this region, of 92.31%, having the exact same values of TP and FN. Comparing the TNR of the models, it is clear that the L7c2-UP-D is slightly better than the other models, as it presents a TNR of 72.72% while the other models presented only 50.00%, for the L7c2-UP-DM, and 58.33% for the models including upwelling variables. This is clearly evidenced in TableC.6 with the L7c2-D model having the highest F1 Macro score. The best results obtained by the L7c2-UP-D model for the negative class can be related to its reduced number of samples, as this model uses a slightly bigger window size than the other models. Nevertheless, it is clear that, similarly to the L1 Carreço and L5b regions, the simpler model obtains the best result for this zone. The decrease in performance with the addition of meteorological variables, although significant for the negative class samples, is not as prejudicial as in these other zones.

- For the L2 Leça da Palmeira region, it is possible to observe that the models containing meteorological variables present a very similar performance when compared to the other models. The L2-UP-D model correctly classifies all the negative samples achieving a TNR of 100.00%, while only obtaining one FN resulting in a TPR of 92.86%. The three other models, maintaining the same FN, but with an increased amount of TP, achieved a slightly higher TPR of 93.33%. These other models, however, presented some FP lowering theirs TNR to 77.78%, for the L2-UP-DU and L2-UP-DM, and 88.89% for the L2-UP-DMU. This better performance of the L2-UP-D model for the negative class leads to the highest F1 Macro and balanced accuracy scores, as represented in Table C.4. Nevertheless, the results obtained for this region in all models are considered very good. The increase in performance on this region, particularly for the models including meteorological variables can be related to, either more representative validation sets, more significant meteorological data or a simpler DSP distribution in the test set where the previous two factors do not impact performance significantly.

5.1.2 Support Vector Machine

The analysis of the obtained results for the SVM Classification models will follow the same structure as the previously discussed RF classification models. Starting with the non-upwelling models, their obtained results for the L1 Carreço region are presented as confusion matrices in Figure 5.12.



(a) L1-D.

(b) L1-DM.

(c) L1-DH.

(d) L1-DMH.

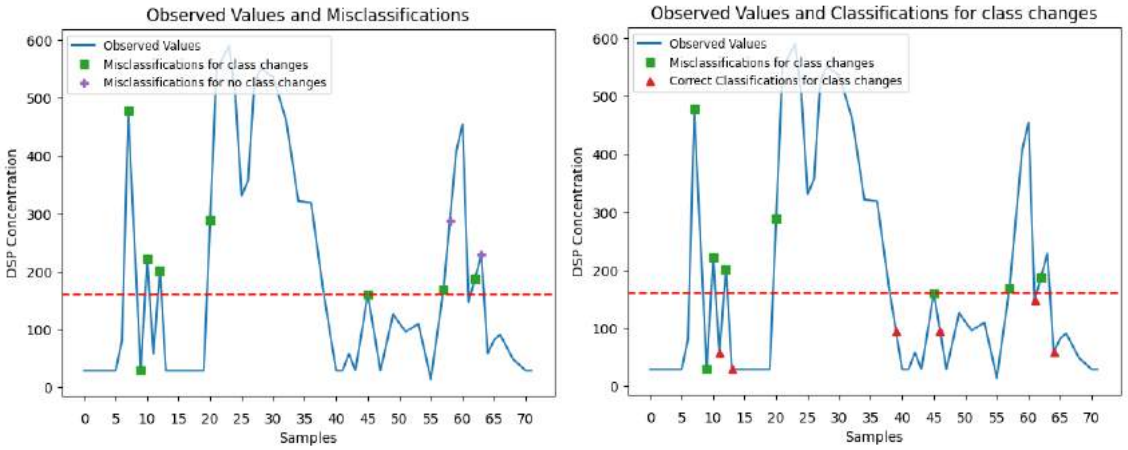
Figure 5.12: L1 Carreço Confusion Matrices for SVM Classification.

- For the L1-D model, it is possible to observe that it presents the highest number of FN and TN while also presenting the lowest amounts of TP and FP. From these values, a TNR of 97.62% and a TPR of 70.0% are obtained.
- An increased amount of FP, while maintaining the same quantity of TP and FN and registering a decrease in TN is obtained from the classifications of the L1-DM model in comparison to the previous model. These changes result in a strong decrease in TNR to 85.17% while maintaining the previous models' TPR. This considerable decrease in performance for the negative class overshadows the model's good TPR resulting in it being considered the worst performing among the four.

- From the classifications obtained from the L1-DH model, it is possible to observe an increase in TPR to 73.33%, resulting from an increase in correct positive classifications that led to a higher amount of TP with a decrease in FN. This model also obtains a decrease in correct classifications on negative samples, lowering its TNR to 95.24%.
- Finally, the best classification performance on the positive class is obtained by the L1-DMH model registering the highest TPR at 76.67% with an increase in TP and decrease in FN. This model, however, fails considerably at correctly classifying negative samples,presenting a lower TNR at 88.10%, only better than the one registered by the L1-DM model.

From this early analysis it is not easy to select the best model with certainty, as the model that obtained the highest TPR also obtained the second lowest TNR and the model that obtained the best TNR obtained only the third best TPR. A promising option seems to be the L1-DH model, that presented a balance between the two metrics, presenting the second highest values for both TPR and TNR.

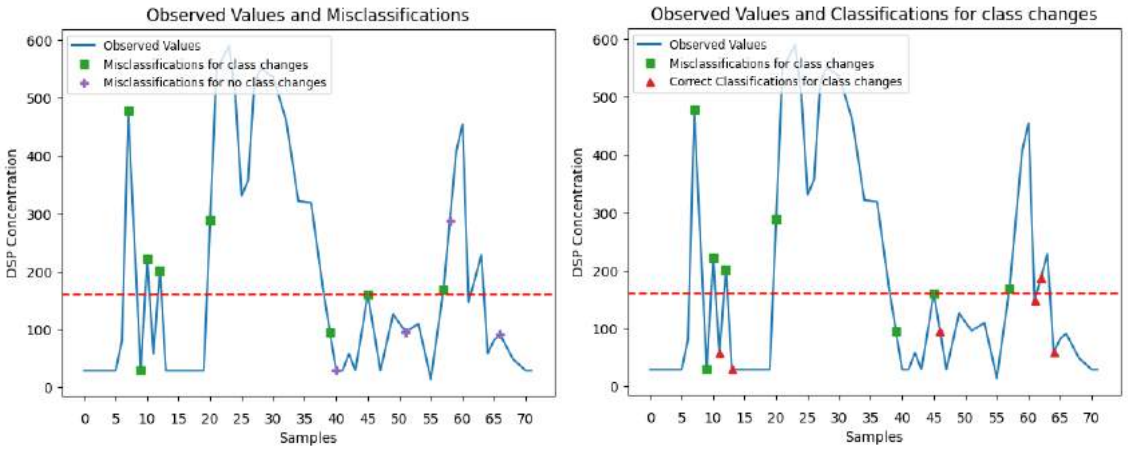
To further explore this hypothesis it is necessary to study the conditions around the misclassifications and the results of the classifications around the contamination threshold, as performed in the analysis of the previously discussed models. Starting with the L1-D, its misclassifications are represented in Figure 5.13(a) with 80% of the misclassifications happening on class changing events. From the displayed graph it is noticeable that the misclassifications follow the patterns previously identified, with the misclassifications for class changes resulting from abrupt variations on DSP concentration or small fluctuations when the values are close to the threshold. As noticed in the previous models, the misclassifications occurring in samples with no class change are always preceded by a misclassification in a class change sample. In Figure 5.13(b), the correct and incorrect classifications for class changing situations are displayed for the same model, with 61.54% of incorrect classifications for class changing samples. From the graph, it is possible to identify that this model can correctly classify 4 drops in DSP concentration, with 2 of them being in the early period of constant variation.



(a) Misclassifications. (b) Class Changes.

Figure 5.13: SVM L1-D and class changes.

The class change information is also present in Figure 5.14 for the L1-DMH model. From the graph on the left, it is possible to observe that, although the percentage of misclassifications of class changing samples is reduced to 66.67%, this is purely motivated by the increase in misclassifications of non-class changing samples, with these misclassifications occurring in the negative class. In the other graph, it is possible to identify that this model also manages to correctly classify the 2 early drops in DSP classification but fails the third one. Despite this, the model manages, contrary to the other models, to correctly predict a positive class on class change, with 61.54% of incorrect classifications as the previous model.

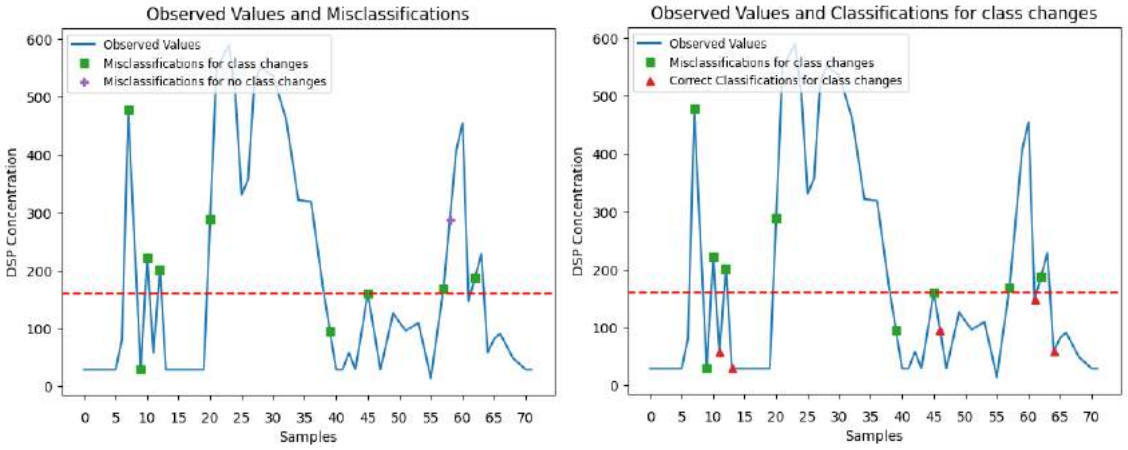


(a) Misclassifications. (b) Class Changes.

Figure 5.14: SVM L1-DMH misclassifications and class changes.

Finally, it is necessary to compare the performance of the L1-DH model on class change

situations to the previous models. In Figure 5.15(a), the misclassifications of this model are categorized. From the graph, it is possible to observe a reduction of misclassifications on non class changing events and an increase in the opposite scenario, resulting in 90% of misclassifications happening in class changing events. This increase in misclassifications in class changing events, implies a reduction of correct classifications on class changing samples, as evidenced in Figure 5.15(b), resulting in 64.29% of misclassifications for class changing samples.

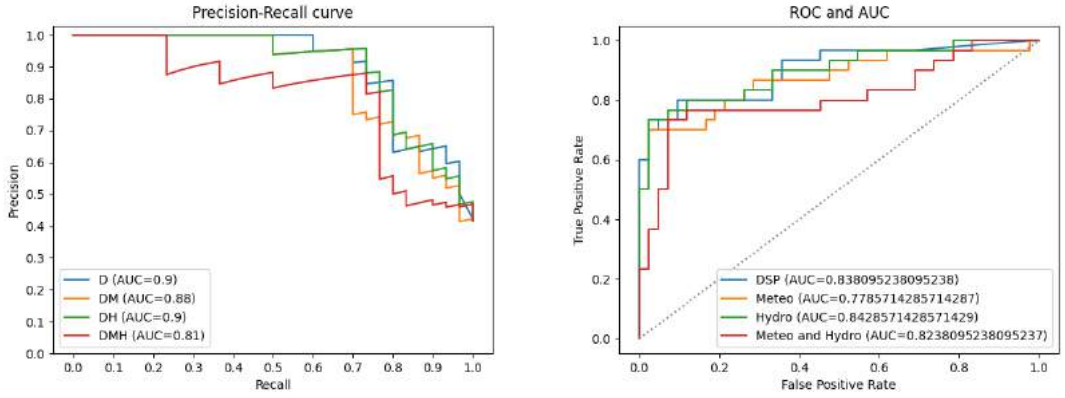


(a) Misclassifications.

(b) Class Changes.

Figure 5.15: SVM L1-DH misclassifications and class changes.

The previous analysis of misclassifications and class changing events for the 3 models being discussed, highlights positive and negative qualities in all of them. For the L1-DMH model, although, it shows an increased classification power on class changing events, as represented in Table 5.3 with the highest recall score, it is counter-balanced by its poor ability to classify non-class changing events. The two remaining models, present a similar performance, with a decision between them relying on a trade-off between power to correctly classify samples on class-changing events, of the L1-D model, and the ability to correctly classify the positive and minority class, that represents a bigger threat in the real world scenario, of the L1-DH model. A slight preference, in the studied scenario, for the L1-DH model is evidenced in Table 5.3, presenting the highest F1 Macro and balanced accuracy scores of all models. As they perform similarly, opting for either one is dependent on the scenario, main goal or major threats with this variability clearly evidenced in Figure 5.16.



(a) PR-REC. (b) AUC-ROC.

Figure 5.16: L1 Carreço SVM Classification curves.

Table 5.3: SVM Classification metrics for L1 Carreço

Set	Metric	Dataset			
		D	DM	DH	DMH
Train	Balanced Accuracy	0.8393	0.8956	0.8741	**0.9417**
	Average Precision	0.9378	0.9657	0.9503	**0.9892**
	F1 Macro	0.8401	0.8959	0.8740	**0.9418**
	Recall	0.7568	0.8683	0.8127	**0.9141**
Validation	Balanced Accuracy	**0.8210**	0.7327	0.7884	0.7454
	Average Precision	**0.8738**	0.8198	0.8278	0.8133
	F1 Macro	**0.8240**	0.7329	0.7802	0.7507
	Recall	**0.7207**	0.6056	0.6720	0.7006
Test	Balanced Accuracy	0.8381	0.7786	**0.8429**	0.8238
	Average Precision	**0.9038**	0.8794	0.8993	0.8130
	F1 Macro	0.8495	0.7822	**0.8519**	0.8268
	Recall	0.7000	0.7000	0.7333	**0.7666**

For the other zones, the models achieved very high performances in the L2 Leça da Palmeira and L5b Caparica regions and slightly poorer results in the RIAV1 Triângulo region. As observed in the L1 Carreço region, all models of each region produced similar results.

- For the L2 Leça da Palmeira region, very similar performances were obtained by the L2-D and L2-DM models registering the same amount of TP and very similar amounts of TN and FP. The L2-DM model shows however, a considerable increase in FN, that lower its TPR to 78.79% when compared to the 86.67% of the L2-D model. This reduction in TPR can be attributed to an overall increased number of positive samples resulting from the lower window size being used. The increased performance of the L2-DM model for negative samples, with a TNR of 89.47%, slightly higher than the L2-D's TNR at 86.84%, is not sufficient to overcome the

decrease in TPR. From these small differences in performance, also observable in the scores in Table C.13, it is possible to conclude that the L2-D is better.

- Similar results were obtained for the L5b Caparica region with both models obtaining very results. The highest TNR is obtained by the L5b-D model at 91.67%, with a TPR of 86.84%. This TPR value is surpassed by the L5b-DM registering 89.47%, with a lower TNR at 89.68%. From these results it is possible to observe that the models have a very similar balanced performance when considering their classification ability for both classes. This balance is clearly evidenced by the almost equal F1 Macro values in Table C.14. The choice between both models rely in the real world scenario being studied, where misclassifications for the positive class are less desirable, making the L5b-DM model preferable. This preference is reinforced by the increased correct classification rate of this model, as shown in Figure C.45(b).

- The models for the RIAV1 Triângulo region achieved lower performances than the two previous zones. The RIAV1-D and RIAV1-DM showed the same classification ability on positive samples with a TPR of 73.17% for both models, also with similar TNR at 86.67% and 83.33%, respectively. For the two other models, RIAV1-DH and RIAV1-DMH, the first achieved a TPR of 73.68% and a TNR of 86.49%, very similar results to the two previous models. For the latter, a poorer classification of positive samples resulted in a TPR of 83.78% with a TNR of 83.78%. From these analysis and with the obtained scores, present in Table C.15, it is possible to conclude that the models perform very similarly, with a slight preference for the RIAV1-DH model due failing less classifications for the positive class, as evidenced in Figure C.50(b).

5.1.2.1 Upwelling Results

For the SVM Classification upwelling models, their results are displayed in Figure 5.17 in the same format as the previous analysis. From the predicted classifications of these models, it is possible to observe that, similarly to the RF Classification models that focused on the upwelling environment, only the L1-UP-D and L1-UP-DU models achieved an acceptable performance with the remaining models completely failing to classify the samples, almost always classifying samples as negative class.

- For the best models, L1-UP-D and L1-UP-DU, a TPR of 83.33% was obtained for the first, with the latter achieving only 69.23%. Both these models achieved a TNR of 100.00%, correctly classifying all negative samples.

- The same TNR was achieved by the L1-UP-DM and L1-UP-DMU models, with the first completely failing to correctly classify positive samples, registering a TPR of 0.00%, while the latter managing a low amount of TP, slightly increasing its TPR to 15.38%.

- Very low TPR values were also obtained by the L1-UP-DH and L1-UP-DHU, at 23.08% and 15.38%, respectively. For the first model, some FP were registered, lowering its TNR to 83.33%, while the latter kept its TNR at 100.00%.

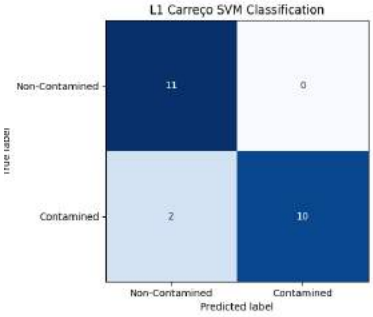
- Finally, for the L1-UP-DMH and L1-UP-DMHU, the former registered a TPR of 0.00%, completely failing to correctly classify the positive class while also presenting a misclassification on the negative class, slightly lowering its TNR to 91.66%. For the latter, a perfect classification on negative samples was achieved, keeping its TNR at 100.00%, for the positive class, however, only one TP was obtained resulting in a TPR of 7.69%.

Following this analysis of the confusion matrices it is relevant to further investigate the negative impact of the upwelling variables when included on the L1-UP-DU when compared to the results of the L1-UP-D model. Comparing the misclassifications of these two models, represented in Figure 5.18, it is possible to observe that both models share the first two misclassifications, with the first occurring in a sample immediately before a class change, showing that both models anticipated too early the decline of DSP contamination to values below the contamination threshold. The latter model, includes however two extra misclassifications on non-class changing samples. The first of these additional misclassifications happens since the model takes longer to adjust to a previous misclassification on a class changing sample, and the in the second, the model, once again, anticipates a change in class from a decrease in DSP contamination, this decrease however does not result in a class change.

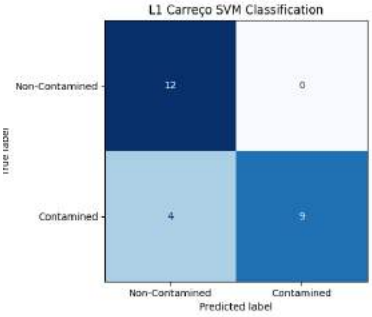
Regarding all the class changing samples, represented in Figure 5.19, it is possible to identify that both models obtain the exact same results on these samples with a correct classification rate of 66.67%.

Focusing on the models that achieved very poor performances, it is possible to observe that, these mostly or even completely predicted a negative class for all samples, failing to correctly classify many of the positive samples, as exemplified in Figure 5.20 for the L1-UP-DHU model. As demonstrated in the graphs, this model shows an initial similar behaviour to the previously analysed models, however, after the first misclassification on a class changing sample and a sharp increase in DSP concentration values, the model is unable to adjust to these changes, failing to correctly classify every following sample.

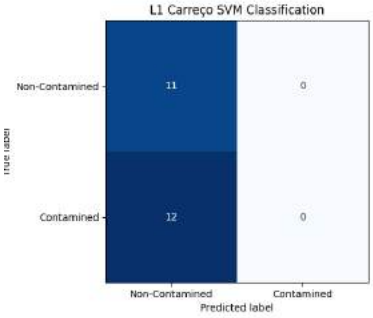
This difficulty in correctly predicting positive class samples aligns with the results obtained for the RF Classification models. The behaviour of these models can be attributed, in part, to unrepresentative validation folds for the L1 Carreço region, with some of them containing very few positive samples. In either type of upwelling models, RF and SVM, it is possible to observe that the inclusion of meteorological or HWP variables correlates to a reduction in performance of the models. The imbalanced data, when combined with these additional variables, that can either be of poor quality or not contributors of knowledge, make the models unable to adjust and predict the positive class. From the PR



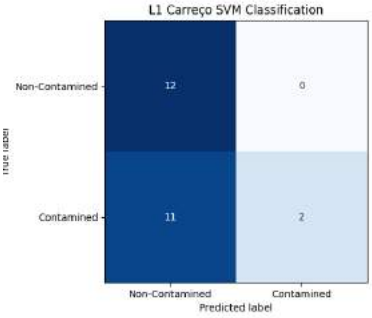
(a) L1-UP-D.



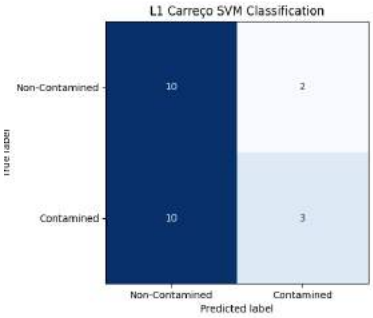
(b) L1-UP-DU.



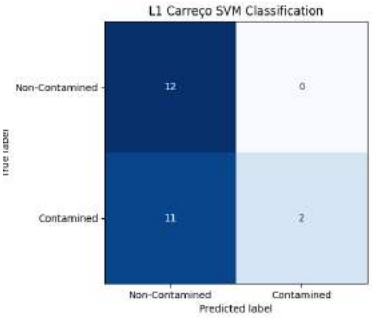
(c) L1-UP-DM.



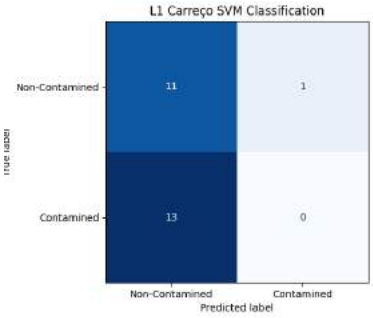
(d) L1-UP-DMU.



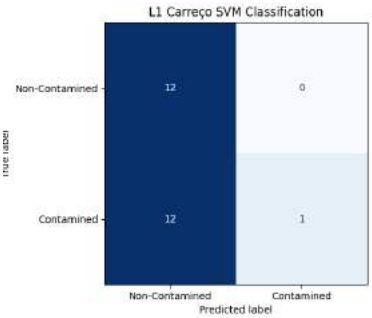
(e) L1-UP-DH.



(f) L1-UP-DHU.

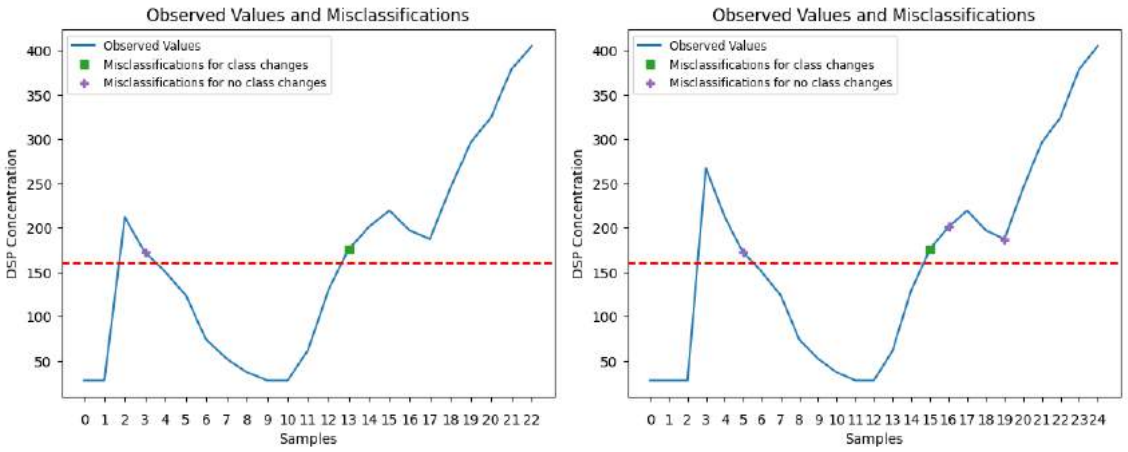


(g) L1-UP-DMH.



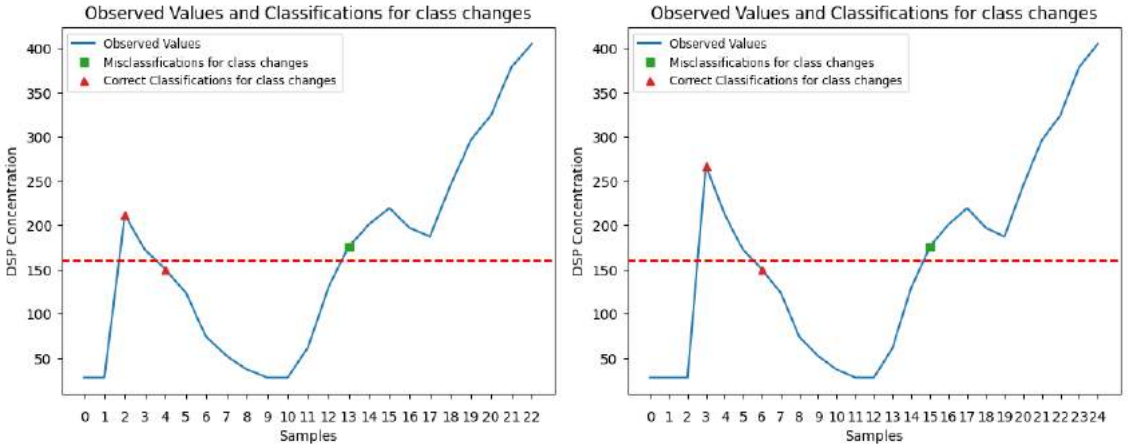
(h) L1-UP-DMHU.

Figure 5.17: Confusion Matrices for SVM Classification upwelling models.



(a) L1-UP-D. (b) L1-UP-DU.

Figure 5.18: SVM L1-UP-D and L1-UP-DU misclassifications.



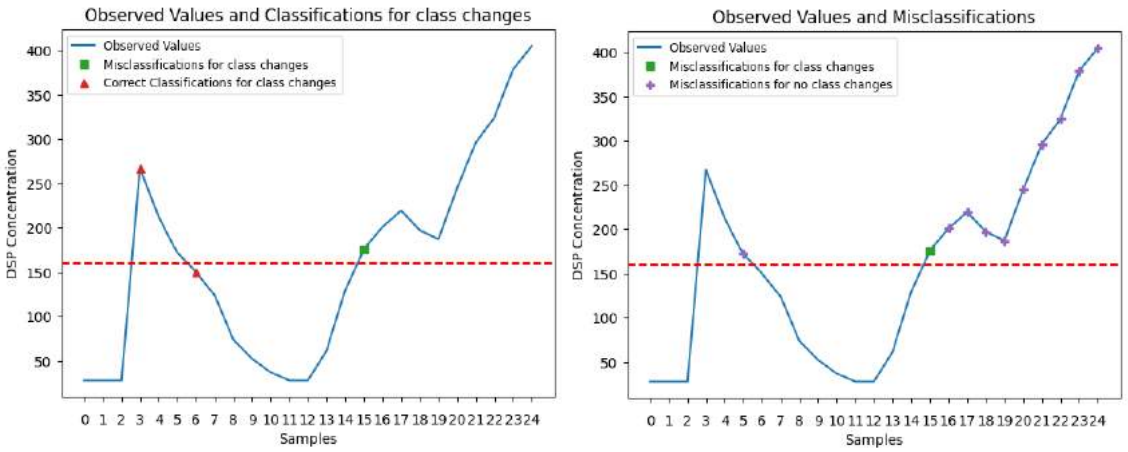
(a) L1-UP-D. (b) L1-UP-DU.

Figure 5.19: SVM L1-UP-D and L1-UP-DU class changing samples.

and AUC-ROC probability curves obtained for these model, displayed in Figure 5.21, it is clear that the L1-UP-D model is clearly more skilled than the other models.

Analysing the scoring metrics in Table 5.4, it is possible to conclude that all models containing meteorological or HWP variables present a steep decrease in performance when compared to the L1-UP-D and L1-UP-DU models. The results for the other zones present similar traits to those found in the previously analysed models.

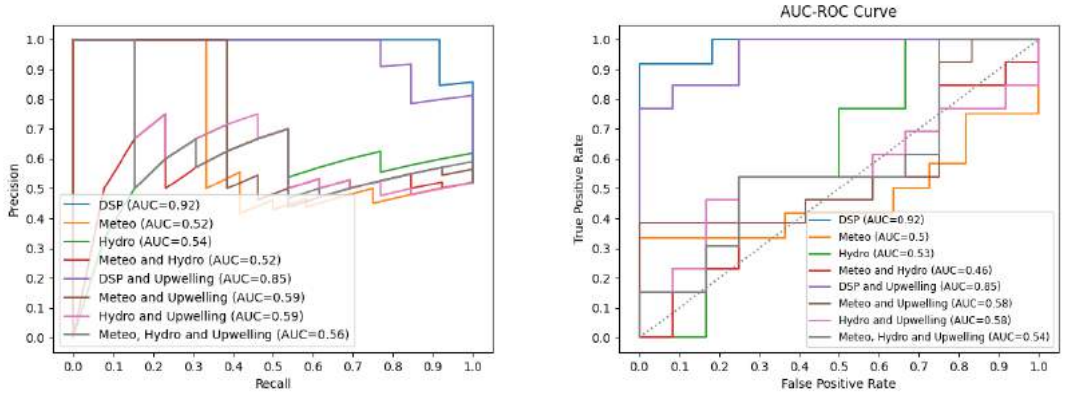
- For the L2 Leça da Palmeira and the L5b Caparica regions, the models containing meteorological data also achieve considerably poorer performances. This decrease in performance is clearly observed in Tables C.10 and C.11 through the F1 Macro and recall scores.



(a) Class changing samples.

(b) Misclassifications.

Figure 5.20: SVM L1-UP-DHU class changes and misclassifications.



(a) PR Curve.

(b) AUC-ROC Curve.

Figure 5.21: L1 Carreço SVM Classification models curves.

- Regarding the L7c2 Porto de Mós region, low but slightly better performances are achieved by these models, this can, however, be justified by a prevalence of positive samples in this zone, with the models struggling, instead, with negative samples. This is evidenced in Table C.12 where F1 Macro and balanced accuracy present lower values than recall.

Table 5.4: SVM Classification Upwelling metrics for L1 Carreço

Set	Metric	Dataset							
		D	DU	DM	DMU	DH	DHU	DMH	DMHU
Train	Balanced Accuracy	0.9374	**1.0000**	**1.0000**	0.9202	0.8583	0.7389	**1.000**	0.9543
	Average Precision	0.9971	**1.0000**	**1.0000**	0.9936	0.9896	0.9753	**1.000**	**1.000**
	F1 Macro	0.9519	**1.0000**	**1.0000**	0.8655	0.7768	0.6118	**1.000**	0.9117
	Recall	0.9833	**1.0000**	**1.0000**	0.8519	0.7165	0.4778	**1.000**	0.9085
Validation	Balanced Accuracy	**0.8762**	0.7880	0.8704	0.8754	0.7927	0.7255	0.8455	0.8497
	Average Precision	0.8935	0.7498	**0.9740**	0.8640	0.8502	0.7438	0.8878	0.8528
	F1 Macro	**0.8899**	0.7646	0.7629	0.7767	0.7517	0.6959	0.7312	0.7407
	Recall	0.8333	0.8413	**1.0000**	0.9412	0.6368	0.5061	0.9804	0.9216
Test	Balanced Accuracy	**0.9167**	0.8462	0.5000	0.5769	0.5321	0.5769	0.4583	0.5385
	Average Precision	**0.9203**	0.8523	0.5217	0.5938	0.5385	0.5938	0.5200	0.5569
	F1 Macro	**0.9129**	0.8377	0.3235	0.4762	0.4792	0.4762	0.3056	0.4048
	Recall	**0.8333**	0.6923	0.0000	0.1538	0.2308	0.1538	0.0000	0.0769

5.1.3 Random Forest versus Support Vector Machine

From the developed classification models analysed previously, it is relevant to compare between the two algorithms used, RF and SVM, in order to evaluate which algorithm performs better and can be more adequate for this approach and the problem being studied. Due to the high number of models developed with different datasets and a focus on different sampling zones that present varying contamination scenarios and complexity degrees, the algorithms will be first compared by region, drawing conclusions in the end.

- For the L1 Carreço region, it is possible to observe that the SVM presents better results than the RF. The SVM classification models for L1-D and L1-DH demonstrate a slightly better performance than the best RF classification models, with SVM L1-DMH performing better than any other model for the positive samples, which, as referred before, are considered the most important in the contamination scenario. Regarding the upwelling models for this region, it is possible to observe that the best model for both algorithms demonstrates the same classification abilities, with it being the simplest model. It is important to notice that the SVM models shows a much steeper decrease in performance with the addition of other variables. This behaviour had already been observed in the non-upwelling models with a considerable decrease in performance for the L1-DM model.

- For the L2 Leça da Palmeira region, the obtained results for both algorithms are very similar with both RF classification models performing slightly better than their SVM counterparts. The results obtained for the upwelling models present the same behaviour as in the L1 Carreço region, with the best models for both algorithms presenting the exact same performance and the SVM models showing a considerable decrease when including meteorological variables.

- Similar results are obtained in the L5b Caparica region for both algorithms. Through the classification scores it is possible to observe that the the RF's L5b-DM ranks slightly higher in terms of overall classification ability. The SVM's L5b-DM, despite ranking in a close second, obtains a higher score for correct classification of positive samples, remaining as a viable option that could be preferable depending on the scenario. For the upwelling models, both algorithms obtain similar poor performances, making neither desirable, with a considerable difference that can only observed for the L5b-UP-DM where the SVM performed much poorly than the RF.
- For the RIAV1 Triângulo region, the results between algorithms are also considered very similar in terms of overall performance. Considering their performance for each class, it is possible to observe that the RF models, overall, perform better for the positive class while the SVM models lead for the negative class.
- Finally, both algorithms show a similar classification power on positive samples for the L7c2 Porto de Mós region. As discussed before, in this region, and contrary to the other zones, the classification models struggle more with negative samples with SVM's L7c2-D presenting the best classification ability for negative samples and overall performance.

From this comparison by region, summarized in Table 5.5, it becomes evident that both algorithms, RF and SVM, commonly obtain very similar performances in their respective models, with both seeming to be viable algorithms for the classification task. Considering the results obtained in these regions, a preference can be attributed to the RF algorithm, since it more frequently slightly outperformed the SVM. The high volatility observed in several SVM classification models when including additional variables strengthens the preference for the other algorithm, as it seems to be more robust.

Table 5.5: Summary table of the classification results

Region	Best Classifier	Best Variable Combination	
		Random Forest	Support Vector Machine
L1 Carreço	Support Vector Machine	D	DH
L2 Leça da Palmeira	Random Forest	D & DM	D
L5b Caparica	Random Forest	DM	DM
RIAV1 Triângulo	Random Forest	DH	DH
L7c2 Porto de Mós	Support Vector Machine	D	D

5.2 Regression

For the regression models, their predicted DSP values will be presented and analysed along with the real DSP values. Following the same structure as the Classification models,

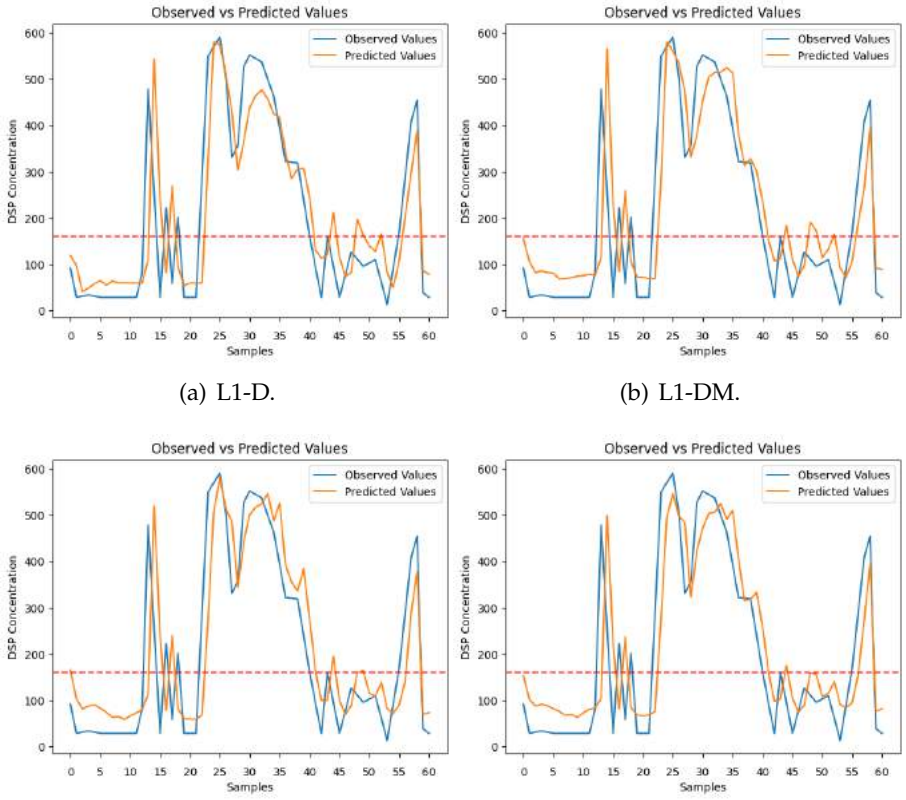
the results for the RF models will be presented first, followed by the SVR results and finally, a comparison between both algorithms.

5.2.1 Random Forest

Starting with the RF, the predictions obtained by the developed models are presented in Figure 5.22, with the real DSP values being represented by the blue line and the predicted ones by the orange line. Although the predicted concentration values seem relatively similar between all models, it is possible to identify some small crucial differences that make some models better and more desirable.

- Starting with the L1-D model, it is possible to observe that it starts by predicting the initial DSP values better than any of the other models but slightly overestimates the first spike in contamination. This is followed by predicting the DSP values very close to the actual value during the second and third peaks with a slight underestimation on the third peak and a considerable overestimation of the following low concentration values, wrongly predicting some of them as contaminated.
- A similar performance was obtained by the L1-DM model, slightly exceeding some of the previous early overestimations and taking longer to adjust to the decrease following the third contamination spike.
- For the the L1-DH model, the concentration values of all peaks were almost precisely followed, however, the decrease following the third spike was very poorly predicted, showing a prediction lag much higher than any of the other models.
- Finally, the L1-DMH model also managed to predicted the contamination of most samples precisely, particularly for the samples with low DSP concentrations between samples 40 and 55. Similarly to the previous model, the L1-DMH shows a considerable lag in its predictions.

For all models, it is possible to observe in their predictions, a slight deviation, usually corresponding to a week of delay. However, for the major peaks in concentration, this gap is slightly sharper in the L1-DM and L1-DMH models. Analysing the feature importances for these models, presented in Figure 5.23, shows that the DSP of the previous week is, by far, the most influential feature, completely dominating the others. This dominance and the delay in predictions imply that the models are very dependent on this variable, heavily basing their predictions on the previous DSP value.



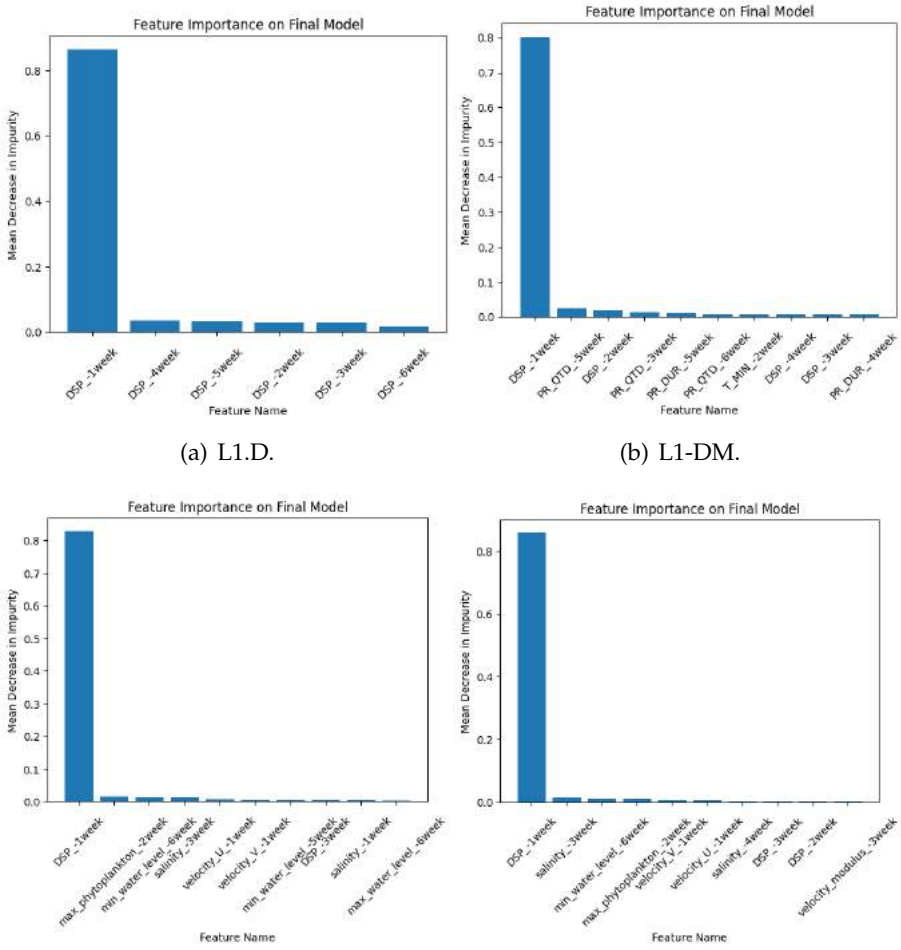
(a) L1-D.

(b) L1-DM.

(c) L1-DH.

(d) L1-DMH.

Figure 5.22: RF Regression predictions for L1 Carreço.



(a) L1.D. (b) L1-DM.

(c) L1-DH. (d) L1-DMH.

Figure 5.23: L1 Carreço RF Regression models feature importance.

Finally, considering the previous observations and the scores in the metrics being used to study and compare these models, present in Table 5.6, it is possible to conclude that the L1-D and L1-DMH models perform better than the others. The former presents the lowest RMSE and MAE, a considerably lower delay and better modelling of the 2 middle peaks while the latter has an increased delay, better predicting, however, concentration values for very small peaks around the contamination threshold.

Table 5.6: RF Regression metrics for L1 Carreço

Set	Metric	Dataset			
		D	DM	DH	DMH
Train	MSE	14670.72	**14019.65**	17080.78	19022.79
	RMSE	119.62	**116.7**	128.53	135.26
	MAE	68.83	**67.39**	75.97	80.88
	R2	0.62	**0.64**	0.56	0.52
Validation	MSE	**14937.33**	15742.20	16817.48	17162.05
	RMSE	**118.32**	120.64	124.92	126.40
	MAE	**81.52**	85.04	89.75	89.62
	R2	**0.45**	0.43	0.39	0.38
Test	MSE	**10452.10**	11309.83	11098.72	10645.18
	RMSE	**102.24**	106.35	105.35	103.18
	MAE	**72.82**	78.21	77.00	76.47
	R2	**0.71**	0.69	0.69	**0.71**

The results obtained for the regions present some similarities to the previous analysis, reinforcing some of the previous observations.

- For the L2 Leça da Palmeira, it is possible to observe that both models, L2-D and L2-DM, obtain several extreme overpredictions. The L2-D manages to predict the high peaks more precisely, but severely overpredicts most small peaks. For the L2-DM, these overpredictions are present but more controlled, being overall more cautious with high concentration predictions, leading to a lower MAE and RMSE than the L2-D model, as displayed in Table C.13.

- The predictions for the L5b Caparica region, show that both models managed to follow the overall DSP variations, however, presented a considerable lag, as observed in other models. Although both models, present similar results, the L5b-D seems preferable, as it contains considerably less overpredictions than in other regions while still precisely predicting the major peaks. The L5b-DM can be slightly less desirable as it is not able to predict the highest peaks as accurately and it presents a slightly higher lag time in some predictions. This small difference between the models, is reflected in their similar MAE and RMSE scores, as evidenced in Table C.14.

- Finally, for the RIAV1 Triângulo region it is possible to observe that the best predictions are obtained by the RIAV1-DMH model as it manages to follow all major

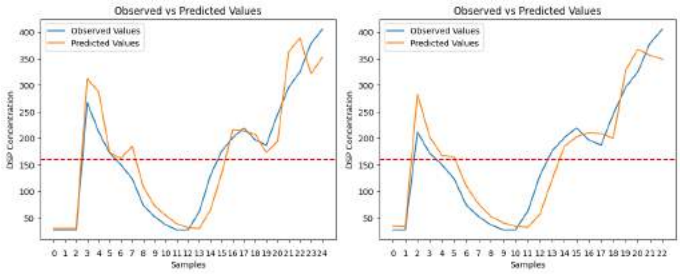
variations in DSP concentration, only considerably underestimating the first peak and the drastic decrease in sample 25, also predicting the final samples very close to the real values. For the RIAV1-D and RIAV1-DH it is possible to identify an additional major peak, resulting from the smaller window size, which is always underestimated with the following peak being almost ignored by both models. making them less desirable options. For the L1-DM model, although similar in many predictions to the RIAV1-DMH model, it shows more overpredictions, particularly between samples 30 and 50, extremely overpredicting the small peak in sample 45. In the scoring metrics present in Table C.15, it is possible to observe the considerably lower RMSE and MAE of the RIAV1-DMH model, confirming that it outperforms the others.

5.2.1.1 Upwelling Results

The predictions for the upwelling RF Regression models for the L1 Carreço region are presented in Figure 5.24. As performed in the previous analysis of this type of models, these predictions will be analysed in pairs of upwelling and non-upwelling variables.

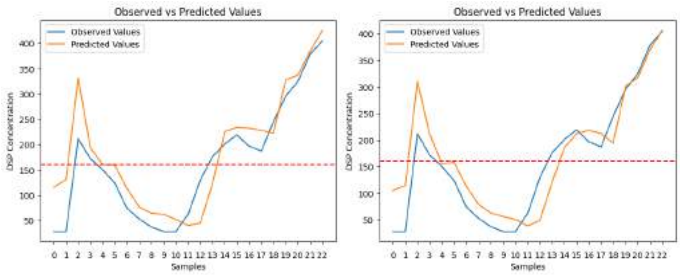
- For the L1-UP-D and L1-UP-DU models, close predictions to the real values are obtained by both models with the latter managing an overall higher performance while being able to better capture the real DSP flow. In contrast, the L1-UP-D model, although still managing close predictions failed to adapt as well to the DSP variation trend showing a prediction series with more pronounced vertices.
- Both models, L1-UP-DM and L1-UP-DMU, overpredicted the first values and their spike in contamination but managed to capture the real DSP values for the last samples almost perfectly predicting their concentrations.
- Similarly to the previous models, the L1-UP-DH and L1-UP-DHU, also overpredicted the first samples, although not as excessively. This improvement, along with good predictions for the remaining samples result in these models being considered better than the previous pair.
- Finally, the L1-UP-DMH and L1-DMHU present very similar results to the previous pair, but registered an increased lag time in their predictions, making them the worst models.

From this analysis it is clear that the predictions between the different models are very similar for the last samples, with significant differences appearing in the first samples. This key difference, highlights the performance of the first pair of models as they excelled in the prediction of these early samples. From the feature importance data for these models, present in Figure 5.25, it is possible to identify that the DSP of the previous week remains the most important variable. In this region, all the RF regression models containing HWP reveal a considerable relevance of the maximum water level of the previous week.



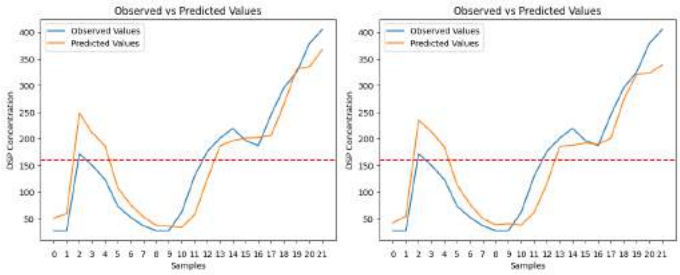
(a) L1-UP-D.

(b) L1-UP-DU.



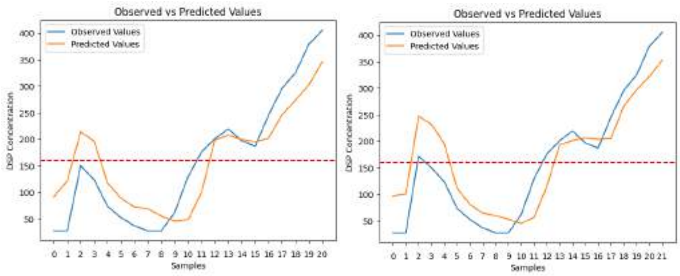
(c) L1-UP-DM.

(d) L1-UP-DMU.



(e) L1-UP-DH.

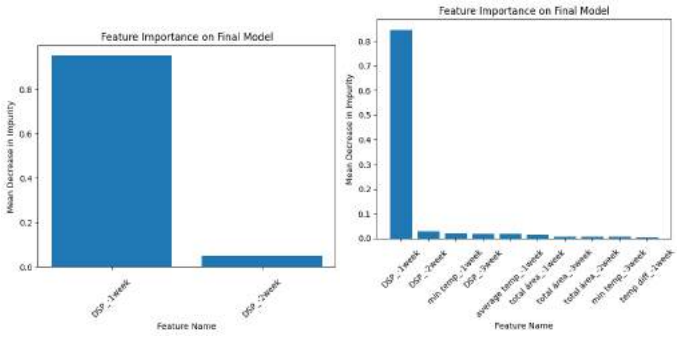
(f) L1-UP-DHU.



(g) L1-UP-DMH.

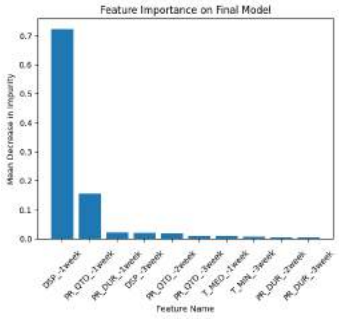
(h) L1-UP-DMHU.

Figure 5.24: RF Regression predictions for L1 Carreço.

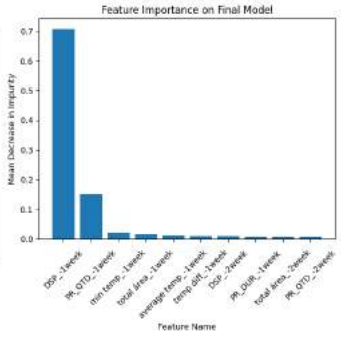


(a) L1-UP-D.

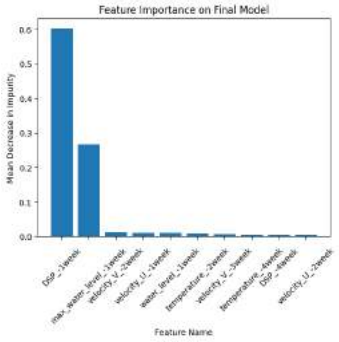
(b) L1-UP-DU.



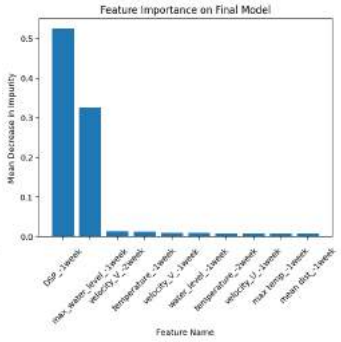
(c) L1-UP-DM.



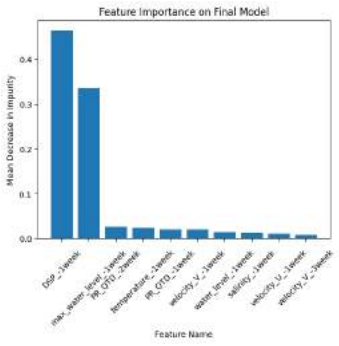
(d) L1-UP-DMU.



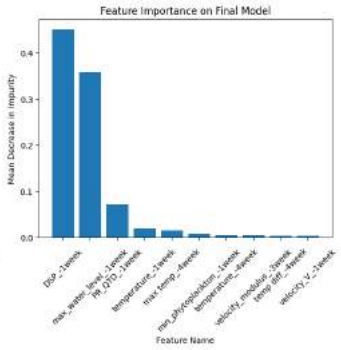
(e) L1-UP-DH.



(f) L1-UP-DHU.



(g) DL1-UP-DMH.



(h) L1-UP-DMHU.

Figure 5.25: L1 Carreço RF Regression models feature importance.

Table 5.7: RF Regression Upwelling metrics for L1 Carreço

Set	Metric	Dataset							
		D	DU	DM	DMU	DH	DHU	DMH	DMHU
Train	MSE	9163.30	6783.65	8509.21	**5558.72**	6589.71	5840.61	8179.71	6481.91
	RMSE	94.43	81.97	90.61	**74.42**	80.34	76.26	88.92	78.87
	MAE	56.10	50.19	54.65	**44.07**	50.56	47.02	58.61	49.09
	R2	0.79	0.83	0.79	**0.86**	0.84	0.85	0.81	0.84
Validation	MSE	**7315.95**	11378.68	19027.00	18494.94	14965.99	14901.42	18236.42	18185.03
	RMSE	**79.70**	99.38	133.44	135.46	121.75	121.23	134.14	134.21
	MAE	**55.14**	73.57	109.71	109.44	99.80	97.06	113.03	112.60
	R2	**0.57**	0.31	-3.17	-2.08	-1.08	-0.84	-1.59	-1.92
Test	MSE	1531.29	**1266.44**	2420.47	1928.77	1489.86	1608.01	2779.55	2339.20
	RMSE	39.13	**35.59**	49.20	43.92	38.60	40.10	52.72	48.37
	MAE	30.57	**30.01**	38.98	34.14	32.42	33.15	45.60	42.29
	R2	0.88	**0.90**	0.81	0.85	0.88	0.88	0.79	0.82

The previous analysis is reinforced by the scoring metrics presented in Table 5.7, confirming that the L1-UP-DU model achieved the best performance, reflected in its lowest MAE and RMSE, followed by the L1-UP-D. The slight decrease in performance for the other models, clearly evidenced in the metrics, is also verified for the other test regions.

- From the results for L5b Caparica region, it is possible to observe that the models struggled with predicting some samples despite the low complexity of the region. For the early samples with constant concentrations values, all models predicted a sudden decreases in DSP in this scenario resulting in the poor performance in this region. The remaining samples are mostly correctly predicted by all models, with the exception of the sudden decrease in the last samples which neither model can accurately predict.
- For the L7c2 Porto de Mós, all models fail to achieve good predictions of the samples, containing constant overpredictions. This behaviour seems to be accentuated for models containing meteorological or HWP variables.
- For the L2 Leça da Palmeira region, very good results were obtained by all models with the L2-UP-D presenting the smallest lag, predicting the peaks almost exactly.

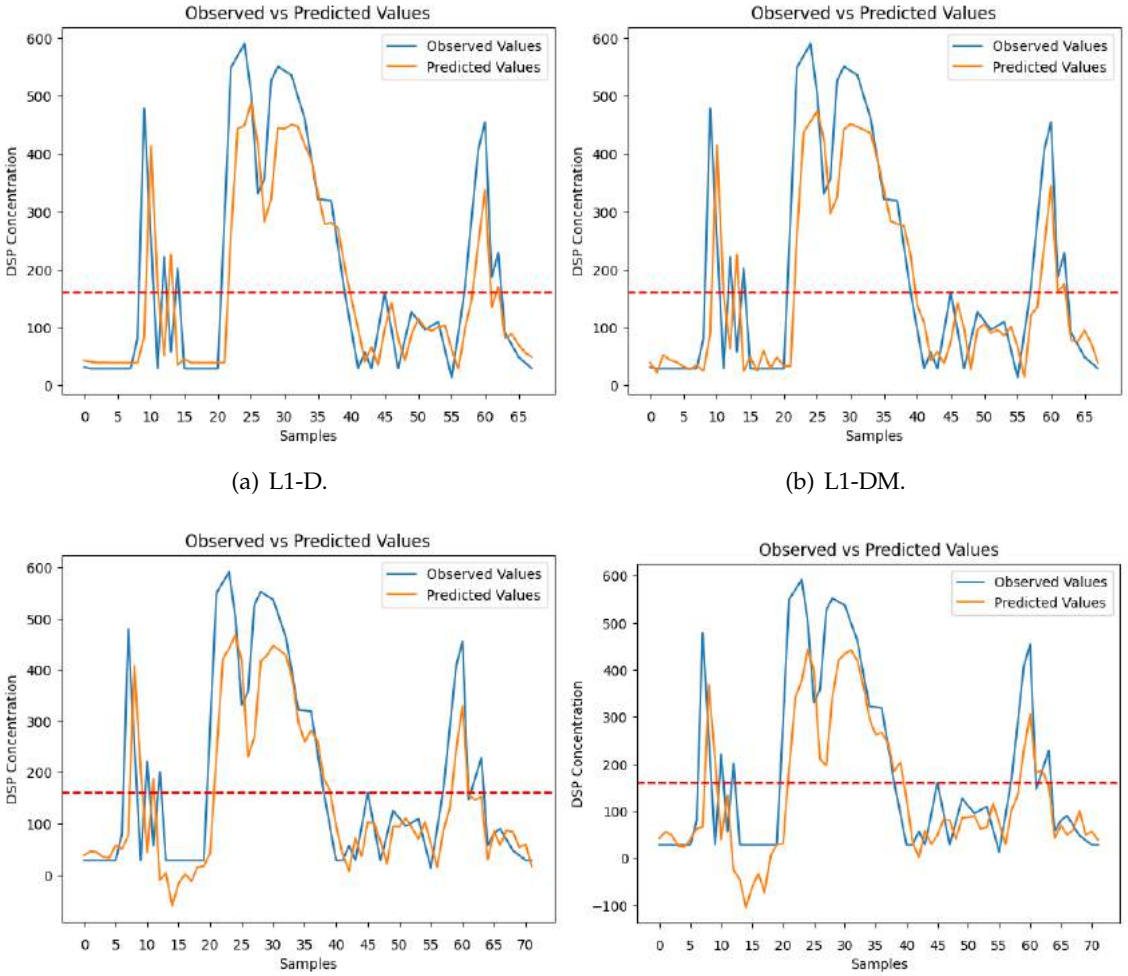
5.2.2 Support Vector Regression

From the results obtained by the SVR models in the L1 Carreço region, displayed in Figure 5.26, it is possible to clearly observe that L1-D performs similar to L1-DM and L1-DH to L1-DMH.

- For the first pair, their predictions seem almost equal, however, the L2-DM presents small fluctuations for groups of samples that contain stable values. Additionally,

both models seem very cautious with their predictions, presenting a constant small lag, meaning that the predictions are heavily based on the most recent DSP values.

- For the other models, although registering similar predictions for most samples to the previous ones, both models predict negative DSP concentration values when faced with a long period of non-detectable contamination and present a very high volatility during the prediction of low values, as observable for the predictions between samples 40 and 55. This discrepancy in performance can be associated with the higher cardinality of the variable sets of these models, leading to lower prediction capability due to more noise in the data.



(a) L1-D.

(b) L1-DM.

(c) L1-DH.

(d) L1-DMH.

Figure 5.26: SVR Regression predictions.

From this analysis, the superiority of the first two models is clear, as evidenced in the scoring metrics in Table 5.8, registering lower MAE and RMSE scores and a higher R^2 than the two other models. The minuscule difference in the scoring metrics, favoring the L1-D model can be associated to the small fluctuations in values observed in the L1-DM model.

Table 5.8: SVR Regression metrics for L1 Carreço

Set	Metric	Dataset			
		D	DM	DH	DMH
Train	MSE	16862.52	16416.94	15855.36	**15728.59**
	RMSE	129.31	127.55	125.50	**124.81**
	MAE	71.80	69.44	70.22	**68.17**
	R2	0.59	0.60	**0.66**	**0.66**
Validation	MSE	**16531.03**	16693.91	17715.40	19474.68
	RMSE	**120.18**	121.13	125.05	132.60
	MAE	**77.85**	80.25	84.12	90.62
	R2	0.53	0.51	**0.55**	0.48
Test	MSE	**9949.69**	10041.46	11067.89	14284.59
	RMSE	**99.75**	100.21	105.20	119.52
	MAE	**66.09**	66.93	71.86	82.12
	R2	**0.71**	0.70	0.66	0.57

For the other zones, it is possible to observe similar patterns in their results to some highlighted in the previous analysis.

- For the L5b-Carreço region, similarly to the L1-DH and L1-DMH, the model containing more variables, L5b-DM, presented a sudden single negative prediction while also registering a high volatility when predicting small concentrations. For the L5b-D model, it is possible to observe that this model bases its predictions completely from the most recent DSP values, with its predictions being an exact representation of the real values with a lag of one week.

- This behaviour, although not as pronounced, is also easily observable for the models in the RIAV1 Triângulo region, particularly RIAV1-D and RIAV1-DM with most predictions resembling the most recent value. For the RIAV1-DMH this phenomenon is not as evident, failing to capture most peaks accurately, This change can be also motivated by the increase in cardinality.

- For the L2 Leça da Palmeira region, this behaviour is still present in the L2-D model but only for a small number of samples. In this region, the models reveal a behaviour similar to the RF Regression upwelling models, with the L2-D overpredicting some samples and the L2-DM predicting the small values better but failing to accurately predict the highest peaks.

5.2.2.1 Upwelling results

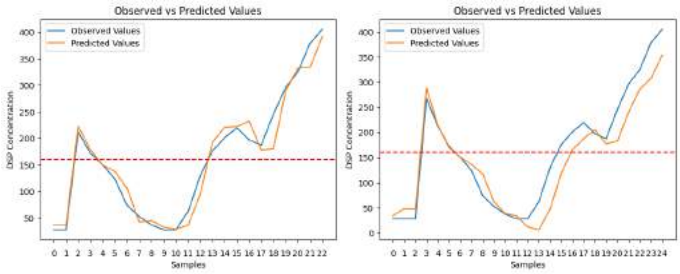
The results obtained for the SVR upwelling models in the L1 Carreço region, displayed in Figure 5.27, reveal a greater performance by the simpler models, almost perfectly predicting the real values for most samples.

- This result is obtained by the L1-UP-D model, that manages to predict DSP concentration very closely to the real values with almost no lag. A similar result is obtained by the L1-UP-DU model, containing however a more significant lag in its predictions.
- For the L1-UP-DM and L1-UP-DMU, most values are also accurately predicted, registering, however, some predicted spikes that do not happen on the real values. Additionally, the latter model has a substantial lag in its predictions.
- For the L1-UP-DH, it is possible to observe that it considerably overpredicts the first DSP spikes but is able to predict the following values with a low lag.
- For the remaining models, the first spike is, overall, well predicted, with the remaining samples being significantly poorly predicted.

The scores present in Table 5.9, emphasise these results, reinforcing the quality of the first model and the poor results achieved by the more complex models.

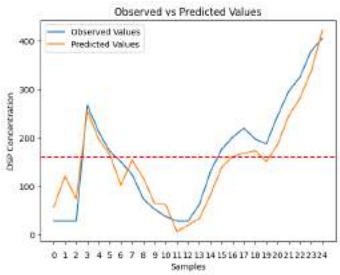
Table 5.9: SVR Regression Upwelling metrics for L1 Carreço

Set	Metric	Dataset							
		D	DU	DM	DMU	DH	DHU	DMH	DMHU
Train	MSE	9509.40	5563.64	8196.53	3602.97	6668.42	3729.97	5218.28	**2593.72**
	RMSE	96.26	71.24	89.72	55.99	81.44	55.49	72.07	**49.21**
	MAE	34.04	33.89	41.20	26.69	32.91	**23.61**	33.17	35.04
	R2	0.77	0.86	0.81	0.91	0.85	0.90	0.88	**0.94**
Validation	MSE	**3480.87**	10994.89	4165.47	10600.99	4420.53	13192.76	5739.88	15446.44
	RMSE	**51.43**	93.44	62.08	100.91	64.60	109.60	74.21	121.12
	MAE	**31.15**	60.09	38.73	61.28	43.26	76.45	51.51	74.94
	R2	**0.84**	0.49	0.56	-0.03	0.62	-0.18	0.29	-0.81
Test	MSE	**523.07**	1423.76	1555.93	4722.97	2377.81	6305.27	4573.40	10791.97
	RMSE	**22.87**	37.73	39.45	68.72	48.76	79.41	67.63	103.88
	MAE	**16.76**	28.78	34.67	57.14	38.72	66.78	56.76	87.03
	R2	**0.96**	0.89	0.88	0.63	0.81	0.50	0.64	0.14

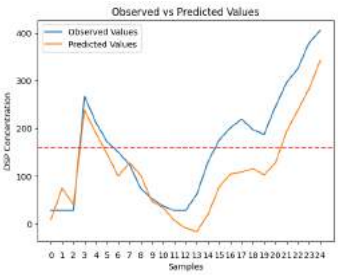


(a) L1-UP-D.

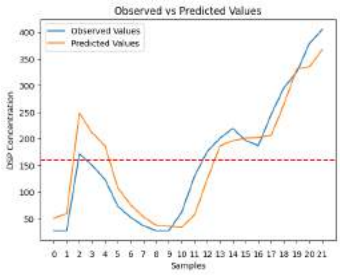
(b) L1-UP-DU.



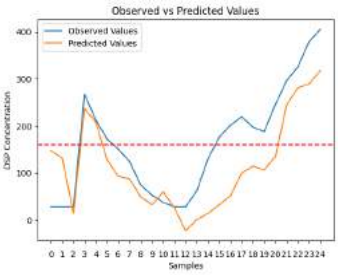
(c) L1-UP-DM.



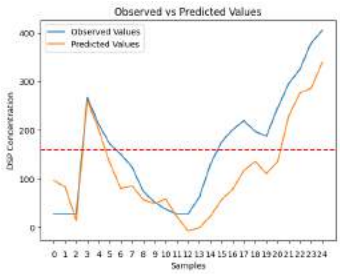
(d) L1-UP-DMU.



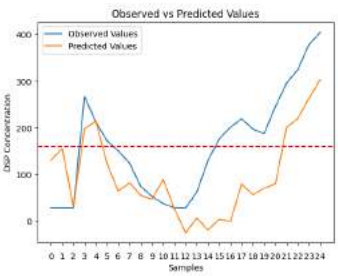
(e) L1-UP-DH.



(f) L1-UP-DHU.



(g) L1-UP-DMH.



(h) L1-UP-DMHU.

Figure 5.27: SVR Regression predictions for L1 Carreço.

The results obtained in the other zones reach varying conclusions, with models in L2 Leça da Palmeira and L5b Caparica containing some of the remarks made for the previous models while the models in L7c2 Porto de Mós showed a different behaviour.

- For the L2 Leça da Palmeira region, all models were capable of predicting the overall DSP trend, with the models containing meteorological or upwelling variables having an increased lag in their predictions.
- This behaviour was also exhibited for the L5b Caparica region, with all models managing to follow the DSP evolution but the models containing meteorological variables showing a considerably higher lag.
- Finally, the aforementioned remarks are not verified for the L7c2 Porto de Mós region, where the models containing upwelling variables obtained the best results. In this region, the simplest model frequently failed to accurately predict the peaks in concentration.

5.2.3 Random Forest versus Support Vector Regression

Similarly to the previous approach, it is important to compare the performance of both algorithms used to develop the models for the regression approach. This comparison will follow the same structure as the previous one, comparing first the regression models by region.

- Comparing both algorithms for the L1 Carreço region, it is possible to observe that the RF achieves better results for the major concentration peaks, failing considerably at predicting the low values. The opposite behaviour is shown by the SVR model, that predicts most lower values very precisely, although with a clear lag, struggling, however, with the major peaks, being able to track them but mostly underpredicting their concentration. In this zone it is also important to consider the poor performance obtained by the SVR when using more variables. For the upwelling models, this pattern is also evidenced, with the simplest SVR model being clearly better than any of the RF regression models. In this region, the SVR models with more variables showed a major decrease in performance while the RF regression models remained more stable.
- For the L2 Leça da Palmeira region, the SVR clearly performs better than the RF regression models, which show an evident tendency to overpredict. In the upwelling models, very similar performances are obtained, with the simplest SVR model showing a slightly better performance with smaller lag. As noted for the previous region, the more complex SVR models considerably dropped in performance.
- The L5b Caparica region presents very ambiguous results, with both RF regression models presenting poorer predictions than the SVR models with a clear lag. For

the SVR models, the L5b-D presents seemingly better results, but a more detailed analysis shows that this model only predicts the previous value for every sample, leading to the predicted values only representing the real values with a lag of one week. The results of this model imply that in reality it is not capable of predicting the DSP concentration for this zone. The SVR's L5b-DM shows more variability in its predictions, with these seeming more independent from the previous value. Nonetheless, similarly to the occurrences in the L1 Carreço region, this model predicts a sudden negative concentration value for one of the last samples. In the upwelling models, as previously analysed, the RF regression models seem to struggle with the first samples but achieve good predictions for the remaining samples with almost no lag. For the SVR models, it is possible to observe that the simplest model obtained the best performance when considering the whole samples, as the models including more variables showed an increased lag.

- For the RIAV1 Triângulo region, both algorithms show very similar performances with several identical predictions. The SVR models seem to perform better when predicting the lower concentration values, heavily basing their predictions on the most recent DSP value. For the RF regression models, a small number of peaks seem to be more precisely predicted and, although, the model overpredicts several samples it predicts some samples with almost no lag. This final remark, emphasizes a preference for the RF regression model in this region.

- Finally, in the L7c2 Porto de Mós region, the SVR models perform better than the RF regression ones. The latter presented some difficulties when predicting the peaks while the former's models that included upwelling variables were able to predict the DSP values more precisely, presenting almost no lag for some of the predictions.

From the previous analysis, summarized in Table 5.10, it is clear that both algorithms show faults in some situations. Considering the comparison of the results of each algorithm, the SVR seems to achieve an overall better performance than the RF. This algorithm, however, presented some anomalies for a few results and also a higher degree of volatility in its predictions, indicating that it requires more caution and attention than the RF regression models, that appeared more stable. Additionally, the SVR mostly obtained its best results from the simplest models, as it frequently showed difficulties when dealing with a variable set with a higher cardinality.

Table 5.10: Summary table of the regression results

Region	Best Classifier	Best Variable Combination	
		Random Forest	Support Vector Regression
L1 Carreço	Support Vector Regression	D	D
L2 Leça da Palmeira	Support Vector Regression	DM	D
L5b Caparica	Support Vector Regression	D	D
RIAV1 Triângulo	Random Forest	DMH	DM
L7c2 Porto de Mós	Support Vector Regression	DM	DU

5.3 Summary

Furthermore, it is necessary to compare both approaches researched in this dissertation, in order to evaluate which can be better suited for the problem. It is important to note that both approaches present major differences and either can be ideal depending on the goals and requirements of the study.

For the classification approach, its main objective was to correctly predict the given samples as contaminated or non-contaminated regarding the DSP max legal concentration, as previously displayed in Table 1.1. The main objective of the regression approach was to accurately predict the real DSP concentration of each sample.

From the obtained results, it was possible to observe that classification approach was mostly successful at achieving its desired goal, with its best models correctly predicting the vast majority of the samples. However, the goal was not completely fulfilled, as the majority of the misclassifications of these classification models happened for the critical class changing samples. These samples were considered the most difficult and crucial, as they represent moments where the contamination switches classes. From the analysis of misclassifications and class changing samples carried out during the analysis of the classification results, it was evident that the vast majority of misclassifications result from class changing samples and that the vast majority of class changing samples are misclassified. For the regression models, the desired goal was not reached, with the developed regression models containing a considerable lag in their predictions and frequently overpredicting or underpredicting the real DSP values.

Both approaches showed no significant variations in their performance throughout the different tested regions. For the classification approach, the developed models seemed to struggle with predicting the samples from the minority class in each region's test set, however, in the RIAV1 Triângulo region, the only one that contained a slight positive majority, not all models struggled with the minority negative class. Additionally, this pattern was not verified by some upwelling models, that, due to their reduced number of samples, commonly present inconsistent behaviour. In the L5b Caparica region, the classification RF models using upwelling datasets presented lower scores on the metrics than most other models, these were, however, skewed due to the very small amount of

negative samples, which the models failed to correctly predict, heavily impacting some of the scoring metrics. For the regression approach, the lag in the predictions was observable for all regions and it was possible to identify regions where the regression models tended to over and underpredict. For the L1 Carreço and L5b Caparica regions, the regression models presented a few small underpredictions towards the highest peaks. In the RIAV1 Triângulo, a reduced number of extreme over and underpredictions could be observed in a few models. The highest tendency for over and underpredictions was identified in the L2 Leça da Palmeira region, where some regression models extremely overpredicted small peaks while other moderately underpredicted the highest peaks.

Considering the different sets of variables tested, namely, DSP, meteorological, HWP and upwelling, similar results were commonly observed between the different combinations with the best one varying throughout approaches and regions. In the classification approach, the use of only DSP or DSP and HWP variables achieved the best results for the L1 Carreço and RIAV1 Triângulo. For the L2 Leça da Palmeira and L5b Caparica region, the use of only DSP or DSP and meteorological variables combinations obtained similarly good results and were considered the best, as they were also the only ones available. For the classification upwelling models, using only DSP clearly achieved the best results. The regression approach results indicate a preference for meteorological variables over HWP variables, since the only model that obtained the best result using HWP also contained meteorological variables. In this approach, using only DSP proved to be the best for the L1 Carreço and L5b Caparica regions while the use of only DSP or DSP and meteorological variables proved to achieve the best results for L2 Leça da Palmeira. For the RIAV1 Triângulo region, the variable combinations using meteorological variables achieved the best results for the regression approach. In the upwelling models of this approach, the use of only DSP or DSP and upwelling variables obtained the best results, with the exception of the L7c2 Porto de Mós region where the DSP and meteorological variables outperformed the use of only DSP.

For both approaches, the use of the upwelling time frame, despite obtaining good results in the simpler models for the reduced sample size, presented several challenges regarding class imbalance and high dimensionality of the datasets, leading to poor performances in many models. A summary of this comparison between the classification and regression approaches is presented in Table 5.11.

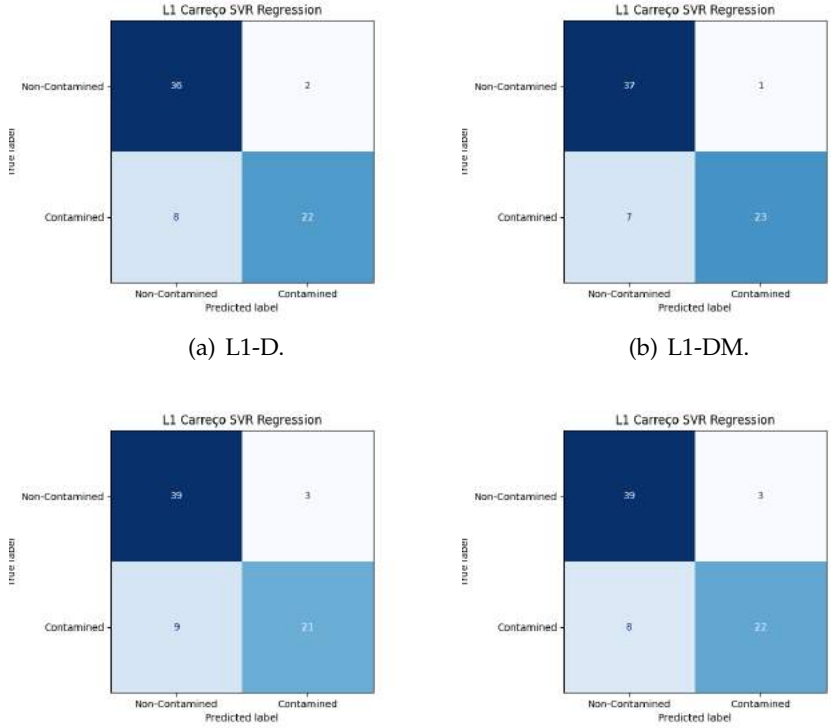
Due to the different nature of the approaches and to allow for a direct comparison between them, the results obtained by the regression approach were converted to the contamination classes and displayed as confusion matrices. For this analysis, the results of the best algorithm for each approach and zone were compared, according to the previous analysis in Sections 5.1.3 and 5.2.3. For the L1 Carreço, the best algorithm for classification was the SVM, while for regression it was the SVR. The SVR converted regression results are presented in Figure 5.28. Comparing these matrices to the ones already analysed for the SVM, in Figure 5.12, it is possible to observe identical results. Evaluating the best TPR and TNR, it is possible to identify that the SVR's L1-DM has a slight increase in

Table 5.11: Summary table of regression and classification approaches

Aspect	Classification	Regression
Main Goal	Predict contamination samples as contaminated or non-contaminated.	Predict the DSP concentration level.
Overall Performance	Best models correctly predicted most of the contamination classes, with the majority of misclassifications occurring on class changing samples.	Consistently presented a significant lag in their predictions, allowing an analysis of the overall evolution of the contamination on the best models but mostly failing at precisely predicting the real DSP concentrations.
Performance across different zones	For all regions, classification models presented good results. For the non-upwelling models, the constructed models generally performed poorer for the minority class in the test set in each region with the exception of the RIAV1 Triângulo region where, despite containing a slight majority of contaminated samples, the SVM models still performed poorer on this class than on the minority non-contaminated samples. This tendency was not easily observable for the upwelling models, as its smaller sample size lead to increased variations in the results.	Similar performances were obtained in all regions always containing a significant lag in the predictions. Models in L1 Carreço and L5b Caparica presented a few moderate underpredictions while models in L2 Leça da Palmeira presented an higher number of both overpredictions and underpredictions. For some models in RIAV1 Triângulo, although in a reduced number, some significant underpredictions or overpredictions could be observed.
Performance using different variable combinations	Similar performances were obtained when using the different sets of variables with best results for the L1 Carreço and RIAV1 Triângulo being mostly obtained when using only DSP or DSP and HWP. For the L2 Leça da Palmeira, only DSP and DSP and Meteorological variables obtained equally good results while DSP and Meteorological presented the best results for L5b Caparica. For the upwelling models, the use of only DSP proved to obtain the best results.	A similar performance between the models was obtained using different datasets, with the use of DSP commonly obtaining slightly better results. For the L2 Leça da Palmeira and RIAV1 Triângulo the models using datasets containing meteorological variables mostly obtained the best results. For the upwelling models, the use of DSP or DSP and upwelling variables proved to achieve the best results.
Performance using upwelling time frames	The models using fewer variables were able to correctly predict the few available samples. An increase in the number of variables frequently led to a steep decrease in performance.	The models were able to predict the general evolution of the DSP values maintaining a consistent lag in most predictions and registering a moderate decrease in performance from the increase in number of variables.

performance when compared to the SVM models as it achieves a TPR of 76.67% and a TNR of 97.37%, values that are not achieved simultaneously by any SVM model.

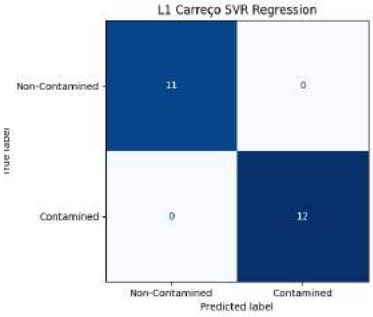
For the upwelling in the L1 Carreço region, the best models for each approach were the RF for Classification and the SVR for Regression. In Figure 5.29, the converted regression results are presented in confusion matrices. Comparing these matrices to the RF matrices, in Figure 5.6, it is possible to observe that similar results were obtained for the best models, with the SVR's L1-D showing a perfect classification with 100.00% for both TPR and TNR. These results represent an increase from the best RF model that achieved a TNR of 100.00% and a TPR of 83.33%. This increase in performance cannot be considered very significant due to the low amount of samples being contemplated.



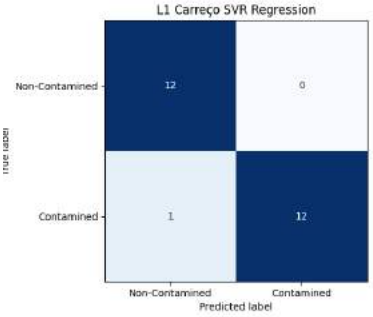
(a) L1-D. (b) L1-DM.

(c) L1-DH. (d) L1-DMH.

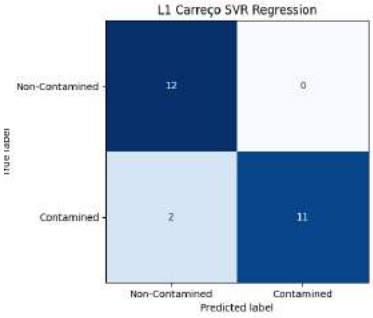
Figure 5.28: L1 Carreço Confusion Matrices for SVR Regression.



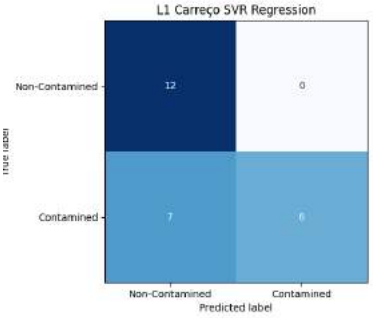
(a) L1-D.



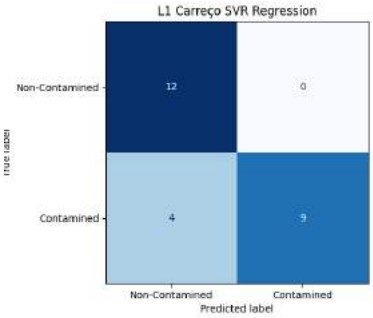
(b) L1-DU.



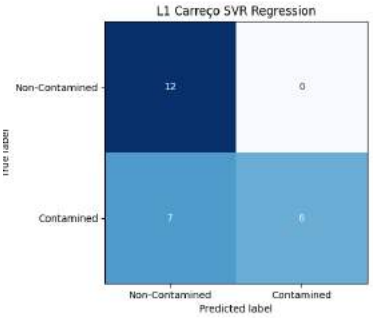
(c) L1-DM.



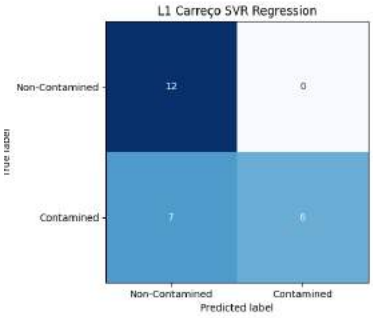
(d) L1-DMU.



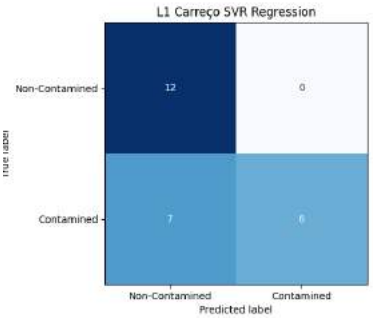
(e) L1-DH.



(f) L1-DHU.



(g) L1-DMHU.



(h) L1-DMHU.

Figure 5.29: L1 Carreço Confusion Matrices for Upwelling SVR Regression.

Comparing these results for the other zones, the same pattern can be found with the best regression models matching the classification results obtained by the best classification models, presenting a small improvement in most situations. Nonetheless, considering that this improvement is not significant, that there is a lack of proper evaluation metrics for the converted results and the concerns raised previously for the SVR Regression models, the proper classification models remain preferable. This preference can be justified by the classification models' robustness and their substantial amount of additional insight on the models performance, including detailed information regarding class changing samples and misclassifications, probability curves and scores on proper evaluation metrics.

6

Conclusions and Future Work

The work developed in this dissertation targeted shellfish contamination by marine biotoxins. To study and better comprehend this phenomenon, the DSP data was used in conjunction with environmental data including meteorological, hydrodynamic and upwelling-related variables. Additionally, the inclusion of this data was meant to evaluate the role of other factors in shellfish contamination. To merge this data, it was necessary to convert all types of data into a common format while carefully treating their missing values. To reach the desired goals, an experimental protocol was constructed to explore different algorithms, namely, RF and SVM, both implemented for classification and regression approaches, using different temporal dimensions and combinations of variables, testing different categories of variables, namely, DSP, meteorological, HWP and upwelling. The protocol was designed taking into account the main challenges: (i) class imbalanced datasets; (ii) distinct spatio-temporal resolution of different data sources of data.

The obtained results allowed a comparison between the explored approaches and algorithms. The classification and regression approaches were compared, following the structure of the summary in Table 5.11, regarding their main goals, overall performance, performance across different zones, using different variable combinations and using upwelling time frame. Both approaches had distinct objectives; for the classification approach, the main goal was to correctly classify the samples into the contamination classes; for the regression approach, the main goal was to accurately predict the contamination level of each sample.

The classification approach managed to largely fulfill its goal, correctly predicting the vast majority of the samples in several models. The objective was not considered as completely achieved as these models struggled in crucial contamination class changing moments. The samples representing these moments were commonly misclassified and also represented the majority of the misclassifications. For the regression approach, its desired objective was not accomplished, as the obtained predictions always presented a significant lag and commonly over and underpredicted the real DSP concentration. Both approaches did not present major variations for the different regions explored. The classification models seemed to struggled at predicting the minority class, commonly

the contaminated class. This behaviour, however, was not verified for some models in RIAV1 Triângulo, the only region with a majority of contaminated samples, and some upwelling models, indicating that this pattern requires further research. For the regression approach, models in the L2 Leça da Palmeira showed an increased tendency for over and underpredictions while in the remaining zones fewer over and underpredictions were registered. Considering the different combinations of variables tested, the classification models presented good results for all variable categories available depending on the focused region, particularly when using only DSP or DSP and HWP variables, when available. For the L2 Leça da Palmeira and L5b Caparica regions, HWP variables were not available and the classification models achieved good results using only DSP or DSP and meteorological variables. For the regression approach, the use of meteorological variables presented better results than the use of HWP variables, with this approach mostly achieving its best results using only DSP or DSP and meteorological variables. For the upwelling time frame and its associated variables, the simpler classification and regression models were able to achieve their desired goal, with the use of upwelling variables being regarded as beneficial for some models. Additionally, the results of the regression models were converted into the contamination classes, to allow a more direct comparison. Through comparing the best models of each approach, both in a classification format, it was possible to observe that the approaches obtained similar class predictions. The similitude between the converted regression results and the classification models' results, demonstrates that the lag in the regression predictions result in misclassifications for the class changing samples in the converted results. This reinforces the claim, that these samples are the most challenging and complex to correctly predict.

Considering the results obtained using the RF and SVM algorithms, it was possible to observe that, for the best models of classification, the RF and SVM presented similar results in the classification approach, slightly better for the first algorithm. For the best regression models, that were concluded to be ineffective, the SVR seemed to outperform the RF. This technical superiority was mostly justified by an overprediction tendency of the RF while the SVR consistently completely based its predictions on the most recent DSP value. For both approaches, SVM and SVR presented a stronger decrease in their predictive capability when dealing with an increased cardinality in the variable set when compared to the RF. From the previous remarks, it is possible to conclude that the RF was considered preferable over the SVM, slightly outperforming it in most tested scenarios, and was regarded as more robust than the SVM and SVR, being less prone to strong decreases in performance from noisy data and better at handling an increase in the variable set. Additionally, both SVM and SVR were shown to be unreliable, susceptible to noisier data and unable to work with datasets of higher dimensionality.

Regarding the different sets of variables utilized, the datasets containing only DSP proved to, frequently, achieve the best results or close to the best results. Nonetheless, the use of other variables along with DSP proved in, some models, to achieve better results than using only DSP. The inclusion of meteorological variables, or a combination of them

and other variables besides DSP, sometimes generated noise for the models, resulting in a decrease in performance, heavily impacting the SVM and SVR upwelling models. For the HWP and Upwelling variables, some good results were obtained but their scarce availability requires further research into their potential.

From the upwelling temporal dimension, despite obtaining good results, no major benefits were observed from the use of this temporal frame and its domain-specific variables when compared to the classification and regression models constructed with non-upwelling datasets. The available upwelling datasets used for these models were also comprised of significantly fewer data cases, not allowing any conclusion to be drawn with certainty.

6.1 Future Work

The work developed in this dissertation can be further expanded in the future. Some of the proposed research prompts are described below.

- **Acquire more complete and precise data in order to improve the quality of the datasets.** Regarding the meteorological data, collecting data closer to the DSP sampling coordinates would improve its quality and contribution to the DSP predictions. The HWP and Upwelling variables require an increase in the availability of data for more regions and for an extended time period.
- **Test smaller time frames for the predictions.** In the reviewed literature, different prediction time frames were explored, ranging from daily to monthly, with the daily prediction models frequently presenting better results.
- **Experiment with other forms of merging the coastal upwelling data and study its impact on shellfish contamination.** The merge between DSP and upwelling data proved to be complex and to further research the impact of this phenomenon on shellfish contamination other ways of merging the data should be explored.
- **Explore multi-class classification models.** Explore classification models featuring several contamination classes' schemes containing intermediate contamination classes allowing for deeper understanding of the results and risk assessment.

Bibliography

[1] B. M. H. Areas. "Monitoring of Toxin-producing Phytoplankton in Bivalve Mollusc Harvesting Areas Guide to Good Practice: Technical Application". In: (2020).

[2] V. Asnaghi et al. "A novel application of an adaptable modeling approach to the management of toxic microalgal bloom events in coastal areas". In: *Harmful Algae* 63 (2017), pp. 184–192.

[3] V. Asnaghi et al. "Interannual variability in Ostreopsis ovata bloom dynamic along Genoa coast (North-western Mediterranean): a preliminary modeling approach". In: *Cryptogamie, Algologie* 33.2 (2012), pp. 181–189.

[4] A. Bakun. "Coastal upwelling indices, west coast of North America, 1946-71". In: (1973).

[5] S. Basu et al. "Iterative random forests to discover predictive and stable high-order interactions". In: *Proceedings of the National Academy of Sciences* 115.8 (2018), pp. 1943–1948.

[6] G. Bonaccorso. *Machine learning algorithms*. Packt Publishing Ltd, 2017.

[7] B. E. Boser, I. M. Guyon, and V. N. Vapnik. "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.

[8] G. E. Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[9] L. Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[10] F. Campuzano et al. "Coupling watersheds, estuaries and regional ocean through numerical modelling for Western Iberia: A novel methodology". In: *Ocean Dynamics* 66 (2016), pp. 1745–1756.

[11] S. Cardeira et al. "Chlorophyll a and chemical signatures during an upwelling event off the South Portuguese coast (SW Iberia)". In: *Continental Shelf Research* 52 (2013), pp. 133–149.

[12] J. Cervantes et al. "A comprehensive survey on support vector machine classification: Applications, challenges and trends". In: *Neurocomputing* 408 (2020), pp. 189–215.

[13] Y. Cheng et al. "A novel random forest approach to revealing interactions and controls on chlorophyll concentration and bacterial communities during coastal phytoplankton blooms". In: *Scientific reports* 11.1 (2021), pp. 1–11.

[14] S. A. Condie, E. C. Oliver, and G. M. Hallegraeff. "Environmental drivers of unprecedented Alexandrium catenella dinoflagellate blooms off eastern Tasmania, 2012–2018". In: *Harmful Algae* 87 (2019), p. 101628.

[15] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[16] R. C. Cruz et al. "Forecasting biotoxin contamination in mussels across production areas of the Portuguese coast with Artificial Neural Networks". In: *Knowledge-Based Systems* 257 (2022), p. 109895.

[17] R. C. E. G. d. Cruz. "Forecasting shellfish contamination by marine biotoxins based onmultivariate time series". PhD thesis. 2022.

[18] J. Derot, H. Yajima, and S. Jacquet. "Advances in forecasting harmful algal blooms using machine learning models: A case study with Planktothrix rubescens in Lake Geneva". In: *Harmful Algae* 99 (2020), p. 101906.

[19] P. B. Fenberg et al. "Biogeographic structure of the northeastern Pacific rocky intertidal: the role of upwelling and dispersal to drive patterns". In: *Ecography* 38.1 (2015), pp. 83–95.

[20] J. A. Fernandes-Salvador et al. "Current status of forecasting toxic harmful algae for the north-east atlantic shellfish aquaculture industry". In: *Frontiers in Marine Science* 8 (2021), p. 666583.

[21] A. Ferreira et al. "Assessing phytoplankton bloom phenology in upwelling-influenced regions using ocean color remote sensing". In: *Remote Sensing* 13.4 (2021), p. 675.

[22] F. A. Gers, J. Schmidhuber, and F. Cummins. "Learning to forget: Continual prediction with LSTM". In: *Neural computation* 12.10 (2000), pp. 2451–2471.

[23] B. Gokaraju et al. "A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the Gulf of Mexico". In: *IEEE Journal of selected topics in applied earth observations and remote sensing* 4.3 (2011), pp. 710–720.

[24] C. Guallar et al. "Artificial neural network approach to population dynamics of harmful algal blooms in Alfacs Bay (NW Mediterranean): Case studies of Karlodinium and Pseudo-nitzschia". In: *Ecological modelling* 338 (2016), pp. 37–50.

[25] J. Guo, Y. Dong, and J. H. Lee. "A real time data driven algal bloom risk forecast system for mariculture management". In: *Marine Pollution Bulletin* 161 (2020), p. 111731.

[26] G. Hallegraeff. "Harmful algal blooms: a global overview". In: *Manual on harmful marine microalgae* 33 (2003), pp. 1–22.

[27] J. R. Harley et al. "Random forest classification to determine environmental drivers and forecast paralytic shellfish toxins in Southeast Alaska with high temporal resolution". In: *Harmful Algae* 99 (2020), p. 101918.

[28] T. D. Harris and J. L. Graham. "Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset". In: *Lake and reservoir management* 33.1 (2017), pp. 32–48.

[29] J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A k-means clustering algorithm". In: *Journal of the royal statistical society. series c (applied statistics)* 28.1 (1979), pp. 100–108.

[30] S. A. Henson and A. C. Thomas. "Interannual variability in timing of bloom initiation in the California Current System". In: *Journal of Geophysical Research: Oceans* 112.C8 (2007).

[31] G. E. Hinton, S. Osindero, and Y.-W. Teh. "A fast learning algorithm for deep belief nets". In: *Neural computation* 18.7 (2006), pp. 1527–1554.

[32] IOC-UNESCO. *The Harmful Algal Event Database (HAEDAT)*. 2021. URL: https://obis.org (visited on 2023-01-08).

[33] M. Izadi et al. "A Remote Sensing and Machine Learning-Based Approach to Forecast the Onset of Harmful Algal Bloom". In: *Remote Sensing* 13.19 (2021), p. 3863.

[34] M. G. Jacox et al. "Coastal upwelling revisited: Ekman, Bakun, and improved upwelling indices for the US West Coast". In: *Journal of Geophysical Research: Oceans* 123.10 (2018), pp. 7332–7350.

[35] R. Kudela, S. Seeyave, and W. Cochlan. "The role of nutrients in regulation and promotion of harmful algal blooms in upwelling systems". In: *Progress in Oceanography* 85.1-2 (2010), pp. 122–135.

[36] J. H. Lee et al. "Neural network modelling of coastal algal blooms". In: *Ecological Modelling* 159.2-3 (2003), pp. 179–201.

[37] S. Lee and D. Lee. "Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models". In: *International journal of environmental research and public health* 15.7 (2018), p. 1322.

[38] X. Li et al. "Harmful algal blooms prediction with machine learning models in Tolo Harbour". In: *2014 International conference on smart computing*. IEEE. 2014, pp. 245–250.

[39] G. Liang et al. "Combining convolutional neural network with recursive neural network for blood cell image classification". In: *IEEE access* 6 (2018), pp. 36188–36197.

[40] M. Lima, P. Relvas, and A. Barbosa. "Variability patterns and phenology of harmful phytoplankton blooms off southern Portugal: Looking for region-specific environmental drivers and predictors". In: *Harmful Algae* 116 (2022), p. 102254.

[41] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf.

[42] H. R. Maier, G. C. Dandy, and M. D. Burch. "Use of artificial neural networks for modelling cyanobacteria Anabaena spp. in the River Murray, South Australia". In: *Ecological Modelling* 105.2-3 (1998), pp. 257–272.

[43] E. MARETEC (Marine and T. Center). *MOHID Modelling System*. 2023. URL: https://maretec.org/en/models/models1/ (visited on 2023-09-30).

[44] A. G. Martins. "Unsupervised Spatio-Temporal Analysis of Coastal Upwelling from Sea Surface Temperature Images". PhD thesis. 2022.

[45] M. Mateus et al. "An operational model for the West Iberian coast: products and services". In: *Ocean Science* 8.4 (2012), pp. 713–732.

[46] M. Mateus and R. Neves. "Ocean modelling for coastal management". In: *Revista Internacional de Desastres Naturales, Accidentes e Infraestructura Civil* 7 (), p. 1.

[47] N. Meinshausen and G. Ridgeway. "Quantile regression forests." In: *Journal of machine learning research* 7.6 (2006).

[48] M. Moita et al. "Distribution of chlorophyll a and Gymnodinium catenatum associated with coastal upwelling plumes off central Portugal". In: *Acta Oecologica* 24 (2003), S125–S132.

[49] N. Muttil and K.-w. Chau. "Neural network and genetic programming for modelling coastal algal blooms". In: *International Journal of Environment and Pollution* 28.3-4 (2006), pp. 223–238.

[50] J. Nalepa and M. Kawulok. "Selecting training sets for support vector machines: a review". In: *Artificial Intelligence Review* 52.2 (2019), pp. 857–900.

[51] O. B. P. G. NASA Goddard Space Flight Center Ocean Ecology Laboratory. *OceanColor*. 2018. URL: https://oceancolor.gsfc.nasa.gov/ (visited on 2023-01-26).

[52] S. Nascimento, S. Casca, and B. Mirkin. "A seed expanding cluster algorithm for deriving upwelling areas on sea surface temperature images". In: *Computers & Geosciences* 85 (2015), pp. 74–85.

[53] S. Nascimento, S. Mateen, and P. Relvas. "Sequential Self-tuning Clustering for Automatic Delimitation of Coastal Upwelling on SST Images". In: *Intelligent Data Engineering and Automated Learning–IDEAL 2020: 21st International Conference, Guimaraes, Portugal, November 4–6, 2020, Proceedings, Part II 21*. Springer. 2020, pp. 434–443.

[54] S. Nascimento et al. "Core–shell clustering approach for detection and analysis of coastal upwelling". In: *Computers & Geosciences* 179 (2023), p. 105421.

[55] S. Nascimento et al. "Novel Cluster Modeling for the Spatiotemporal Analysis of Coastal Upwelling". In: *Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings*. Springer. 2022, pp. 563–574.

[56] M. Oliveira. *Evaluating the Role of Environmental Variables on Shellfish Biotoxin Contamination via Supervised Learning*. 2023. URL: https://github.com/v1scal/shellfish_results.

[57] H. de Pablo et al. "The influence of the river discharge on residence time, exposure time and integrated water fractions for the tagus estuary (Portugal)". In: *Frontiers in Marine Science* 8 (2022), p. 734814.

[58] S. Palma et al. "Can Pseudo-nitzschia blooms be modeled by coastal upwelling in Lisbon Bay?" In: *Harmful Algae* 9.3 (2010), pp. 294–303.

[59] B. Park and J. K. Bae. "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data". In: *Expert systems with applications* 42.6 (2015), pp. 2928–2934.

[60] D. A. Pisner and D. M. Schnyer. "Support vector machine". In: *Machine learning*. Elsevier, 2020, pp. 101–121.

[61] G. C. Pitcher et al. "Harmful algal blooms in eastern boundary upwelling systems: A GEOHAB Core Research Project". In: *Oceanography* 30.1 (2017), pp. 22–35.

[62] P. Probst, M. N. Wright, and A.-L. Boulesteix. "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9.3 (2019), e1301.

[63] M. Qin, Z. Li, and Z. Du. "Red tide time series forecasting by combining ARIMA and deep belief network". In: *Knowledge-Based Systems* 125 (2017), pp. 39–52.

[64] F. Recknagel et al. "Artificial neural network approach for modelling and prediction of algal blooms". In: *Ecological Modelling* 96.1-3 (1997), pp. 11–28.

[65] F. Recknagel et al. "Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes". In: *Journal of Hydroinformatics* 4.2 (2002), pp. 125–133.

[66] R. Ribeiro and L. Torgo. "A comparative study on predicting algae blooms in Douro River, Portugal". In: *Ecological modelling* 212.1-2 (2008), pp. 86–91.

[67] S. Salcedo-Sanz et al. "Support vector machines in engineering: an overview". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.3 (2014), pp. 234–267.

[68] M. Schonlau and R. Y. Zou. "The random forest algorithm for statistical learning". In: *The Stata Journal* 20.1 (2020), pp. 3–29.

[69] N. O. Service. *National Oceanic and Atmospheric Administration*. 2022. URL: https://oceanservice.noaa.gov/hazards/hab/ (visited on 2023-09-30).

[70] T. J. Smayda. "Turbulence, watermass stratification and harmful algal blooms: an alternative view and frontal zones as "pelagic seed banks"". In: *Harmful Algae* 1.1 (2002), pp. 95–112.

[71] A. Thomas et al. "Comparison of the seasonal and interannual variability of phytoplankton pigment concentrations in the Peru and California Current systems". In: *Journal of Geophysical Research: Oceans* 99.C4 (1994), pp. 7355–7370.

[72] L. Torgo and R. Ribeiro. "Predicting rare extreme values". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2006, pp. 816–820.

[73] V. L. Trainer et al. "The distribution and impacts of harmful algal bloom species in eastern boundary upwelling systems". In: *Progress in oceanography* 85.1-2 (2010), pp. 33–52.

[74] J. Tweddle et al. "Relationships among upwelling, phytoplankton blooms, and phycotoxins in coastal Oregon shellfish". In: *Marine Ecology Progress Series* 405 (2010), pp. 131–145.

[75] L. Velo-Suárez and J. Gutiérrez-Estrada. "Artificial neural network approaches to one-step weekly prediction of Dinophysis acuminata blooms in Huelva (Western Andaluc´ ia, Spain)". In: *Harmful Algae* 6.3 (2007), pp. 361–371.

[76] L. G. Vilas et al. "Support Vector Machine-based method for predicting Pseudo-nitzschia spp. blooms in coastal waters (Galician rias, NW Spain)". In: *Progress in Oceanography* 124 (2014), pp. 66–77.

[77] M. L. Wells et al. "Harmful algal blooms and climate change: Learning from the past and present to forecast the future". In: *Harmful algae* 49 (2015), pp. 68–93.

[78] J. Wen et al. "Harmful algal bloom warning based on machine learning in maritime site monitoring". In: *Knowledge-Based Systems* 245 (2022), p. 108569.

[79] K. T. Wong, J. H. Lee, and P. J. Harrison. "Forecasting of environmental risk maps of coastal algal blooms". In: *Harmful algae* 8.3 (2009), pp. 407–420.

[80] K. T. Wong, J. H. Lee, and I. Hodgkiss. "A simple model for forecast of coastal algal blooms". In: *Estuarine, Coastal and Shelf Science* 74.1-2 (2007), pp. 175–196.

[81] H. Yajima and J. Derot. "Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases". In: *Journal of Hydroinformatics* 20.1 (2018), pp. 206–220.

[82] A. Ziegler and I. R. König. "Mining data with random forests: current options for real-world applications". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.1 (2014), pp. 55–63.

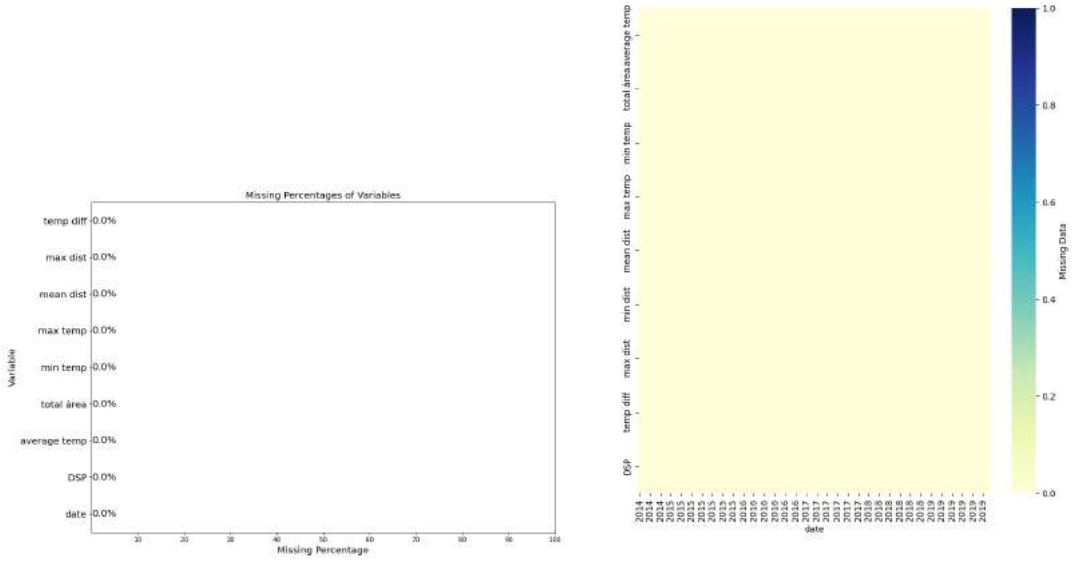
[83] E. Zohdi and M. Abbaspour. “Harmful algal blooms (red tide): a review of causes, impacts and approaches to monitoring and prediction”. In: *International Journal of Environmental Science and Technology* 16 (2019), pp. 1789–1806.

A

Appendix 1: Datasets Construction

A.1 Data Preprocessment

A.1.1 L1 Carreço



(a) Missing percentages.

(b) Missing periods.

Figure A.1: L1 Carreço upwelling missing values.

A.1.2 L2 Leça da Palmeira

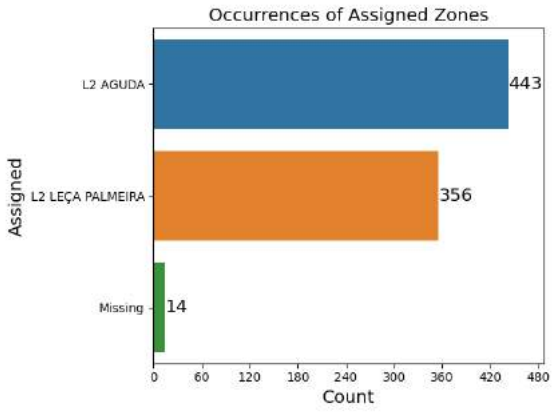
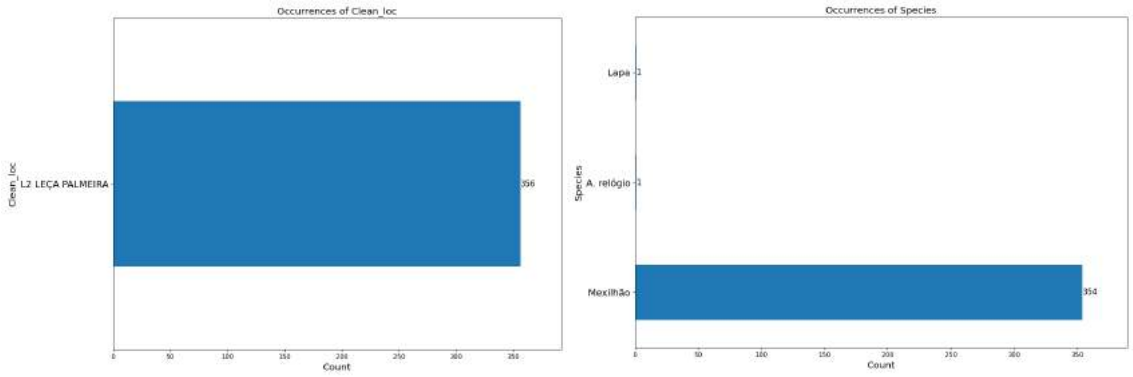


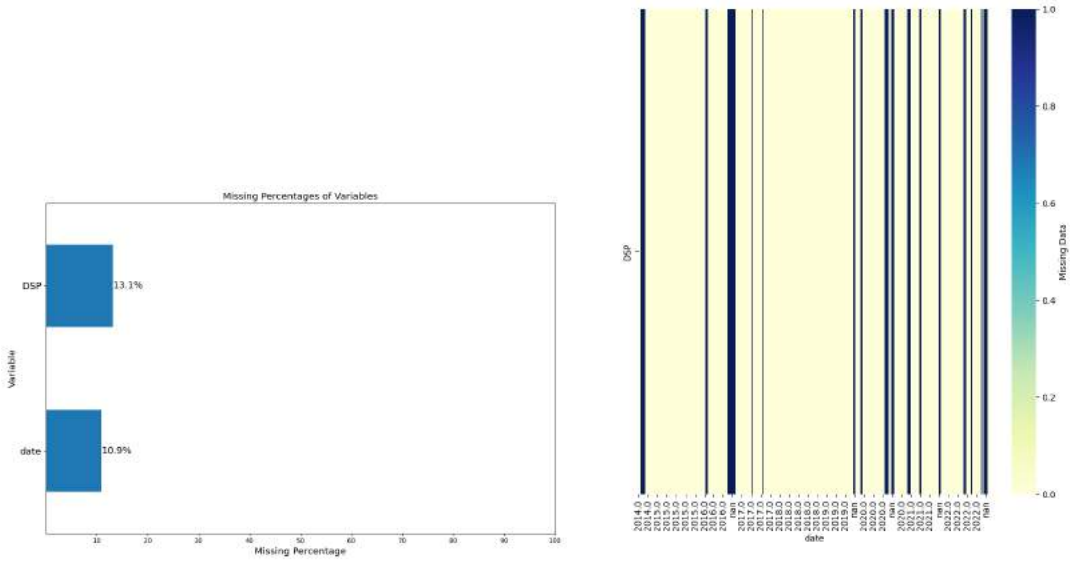
Figure A.2: Occurrences in each assigned sampling zone in the L2 region.



(a) Clean location distribution.

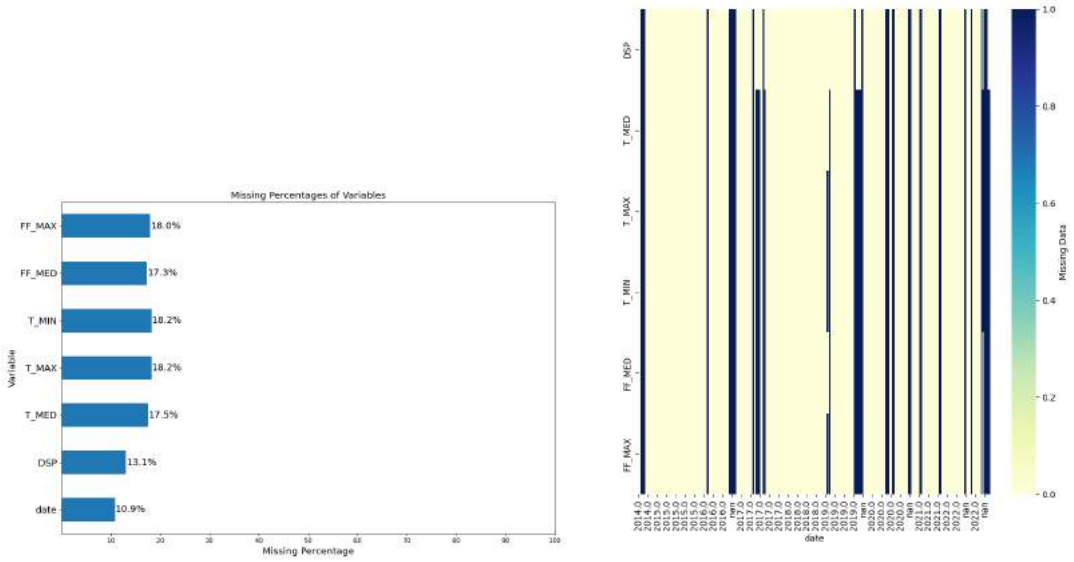
(b) Species distribution.

Figure A.3: L2 Leça da Palmeira clean location and species distributions.



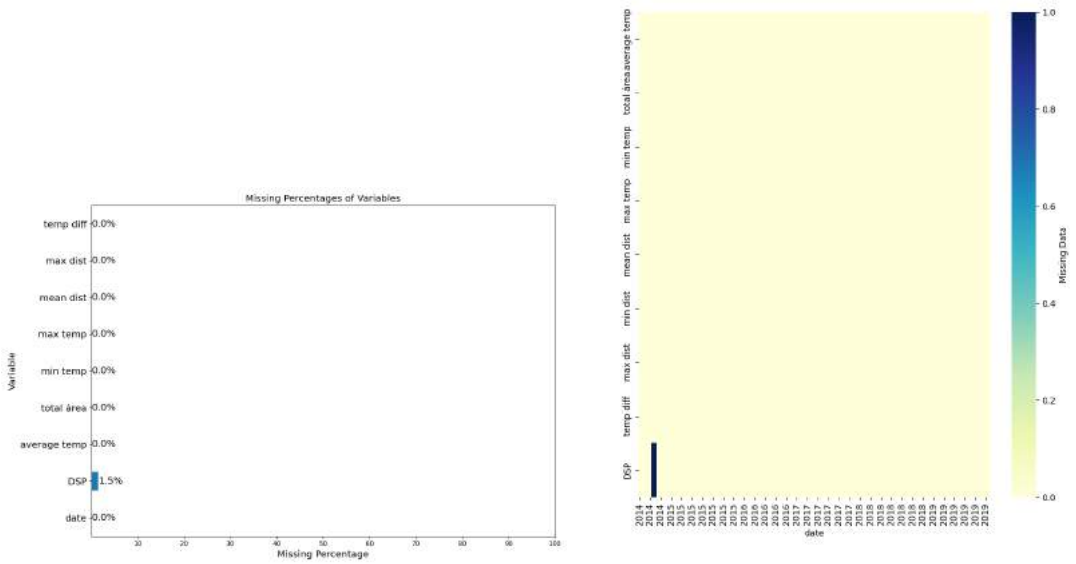
(a) Missing percentages. (b) Missing periods.

Figure A.4: L2 Leça da Palmeira DSP missing values.



(a) Missing percentages. (b) Missing periods.

Figure A.5: L2 Leça da Palmeira meteorological missing values.



(a) Missing percentages.

(b) Missing periods.

Figure A.6: L2 Leça da Palmeira upwelling missing values.

A.1.3 L5b Caparica

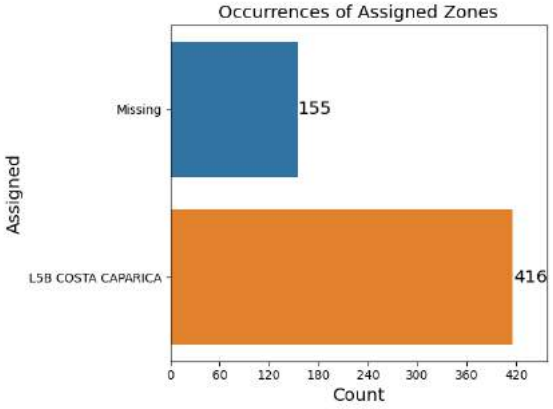
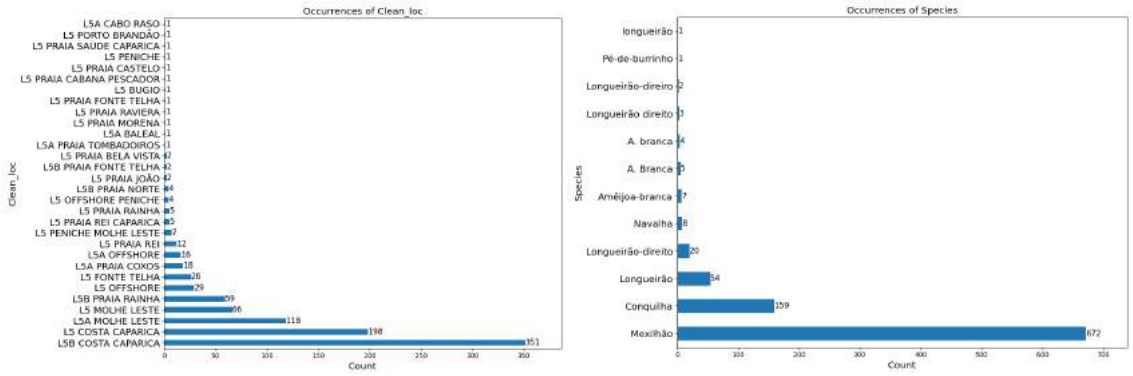


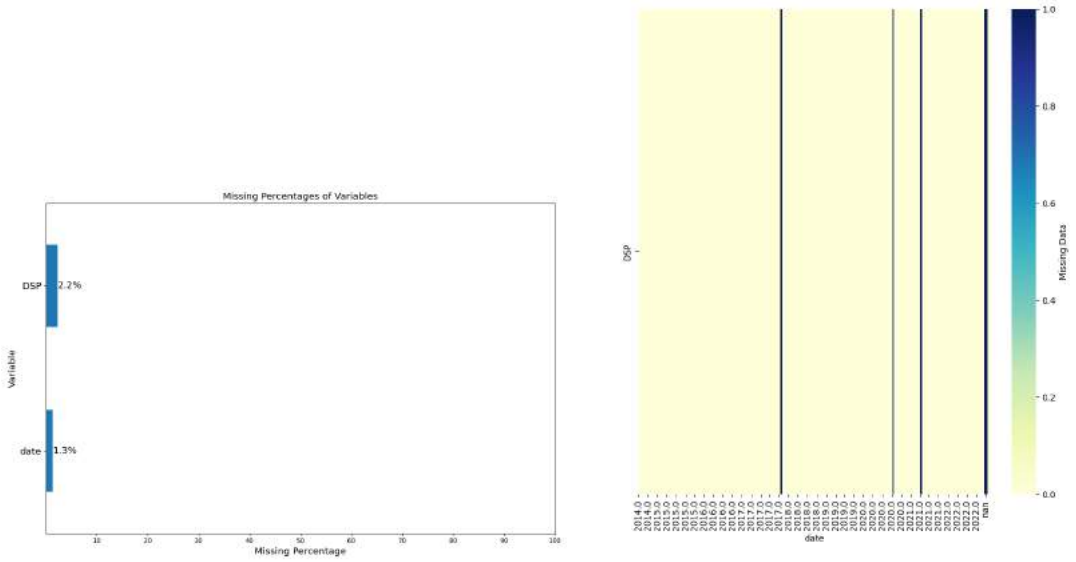
Figure A.7: Occurrences in each assigned sampling zone in the L5b region.



(a) Clean location distribution.

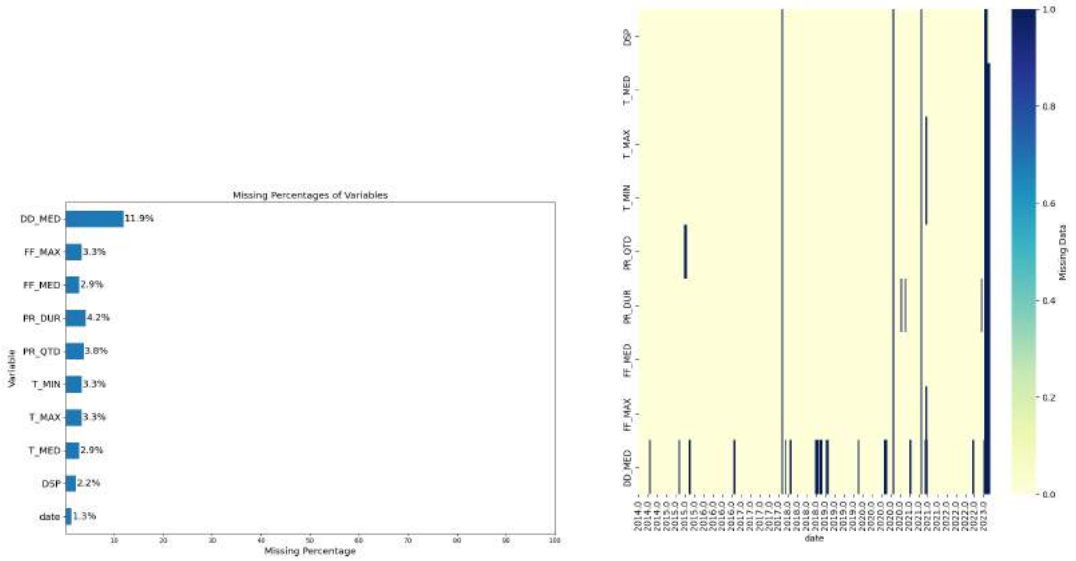
(b) Species distribution.

Figure A.8: L5b Caparica clean location and species distributions.



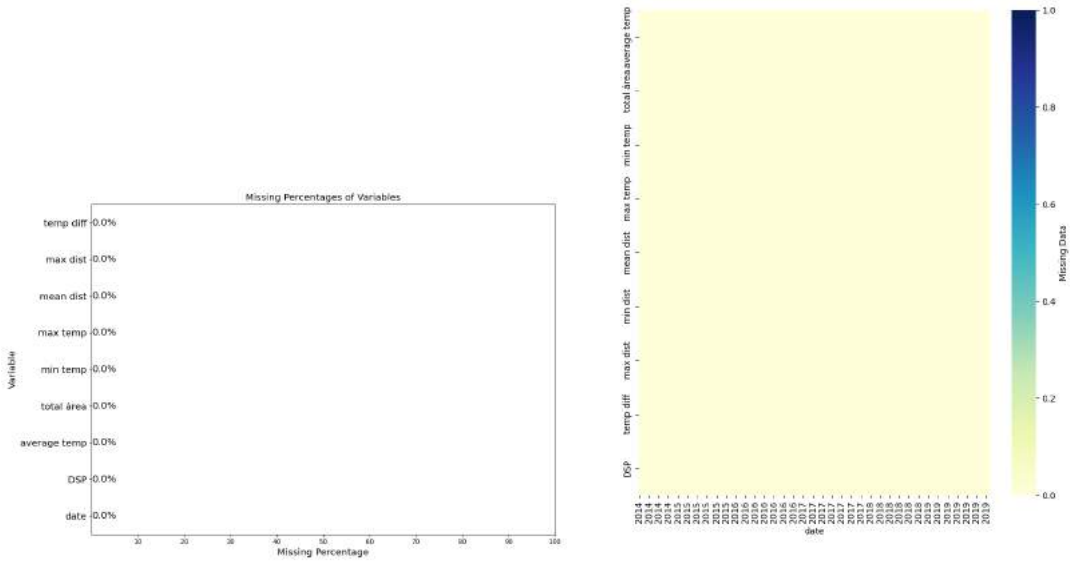
(a) Missing percentages. (b) Missing periods.

Figure A.9: L5b Caparica DSP missing values.



(a) Missing percentages. (b) Missing periods.

Figure A.10: L5b Caparica meteorological missing values.



(a) Missing percentages.

(b) Missing periods.

Figure A.11: L5b Caparica upwelling missing values.

A.1.4 RIAV1 Triângulo

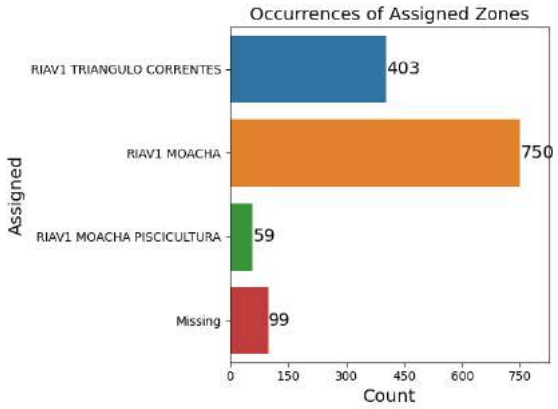
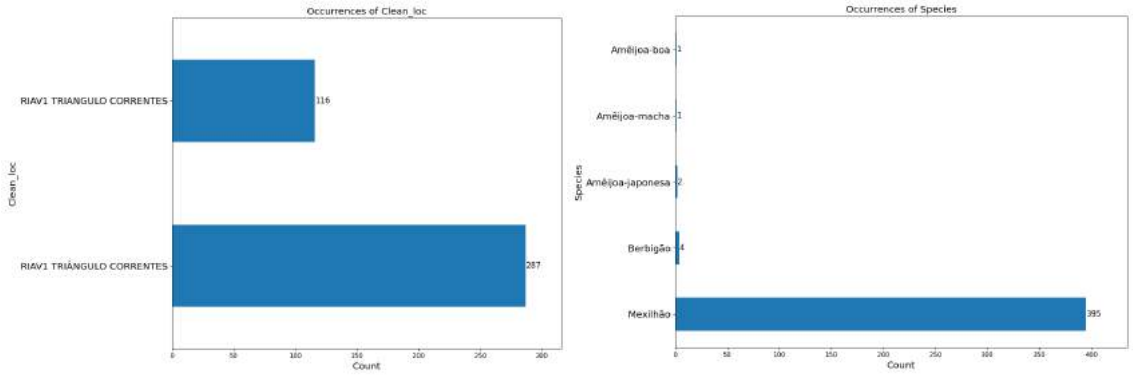


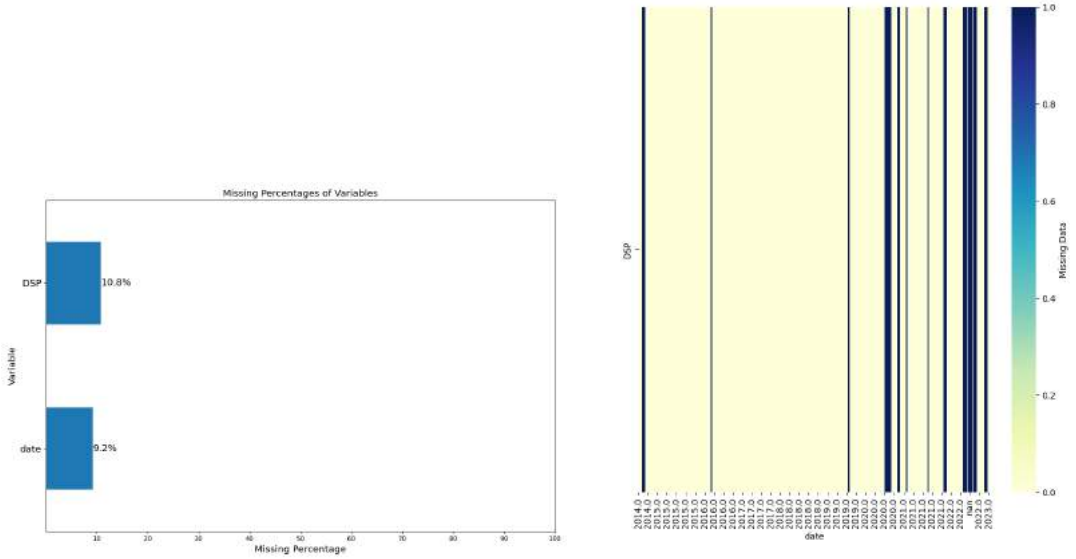
Figure A.12: Occurrences in each assigned sampling zone in the RIAV1 region.



(a) Clean location distribution.

(b) Species distribution.

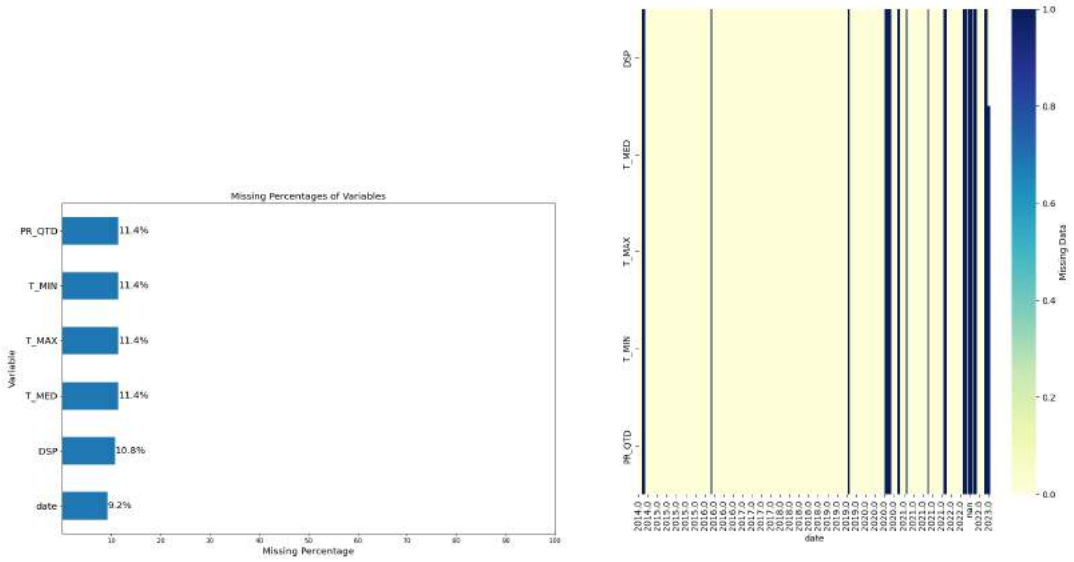
Figure A.13: RIAV1 Triângulo clean location and species distributions.



(a) Missing percentages.

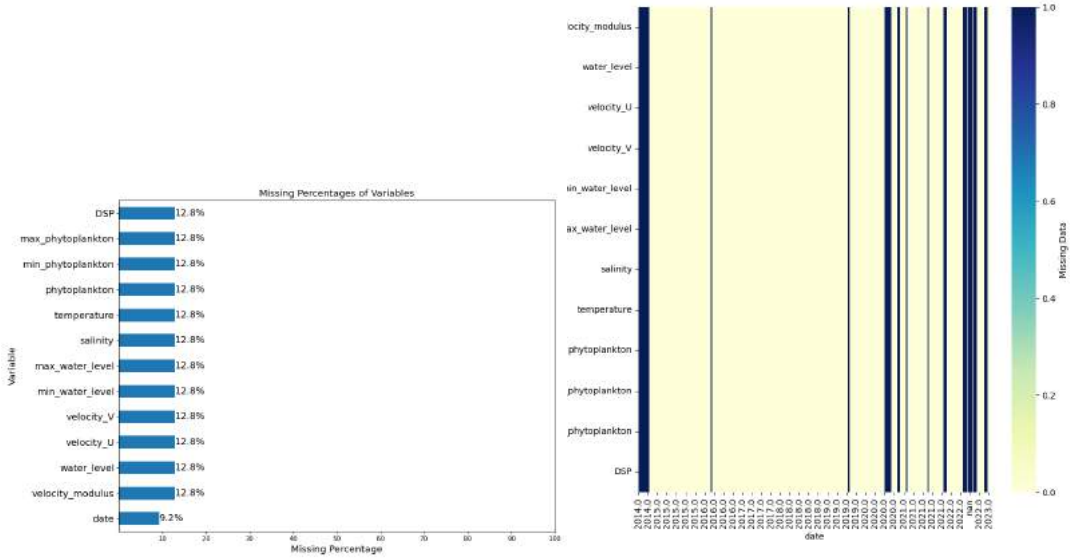
(b) Missing periods.

Figure A.14: RIAV1 Triângulo DSP missing values.



(a) Missing percentages. (b) Missing periods.

Figure A.15: RIAV1 Triângulo meteorological missing values.



(a) Missing percentages. (b) Missing periods.

Figure A.16: RIAV1 Triângulo HWP missing values.

A.1.5 L7c2 Porto de Mós

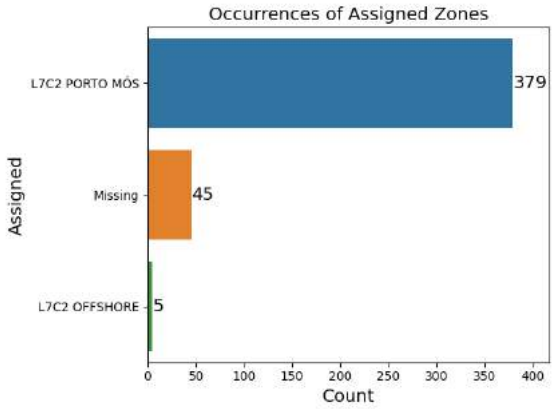
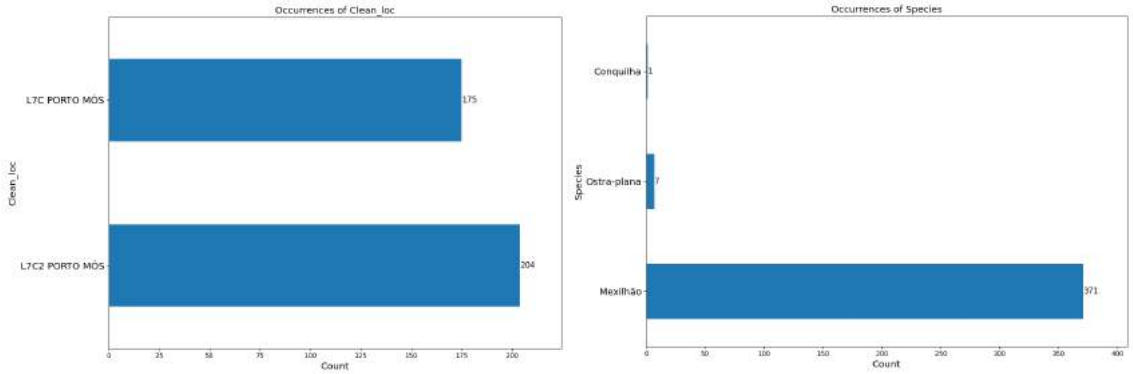


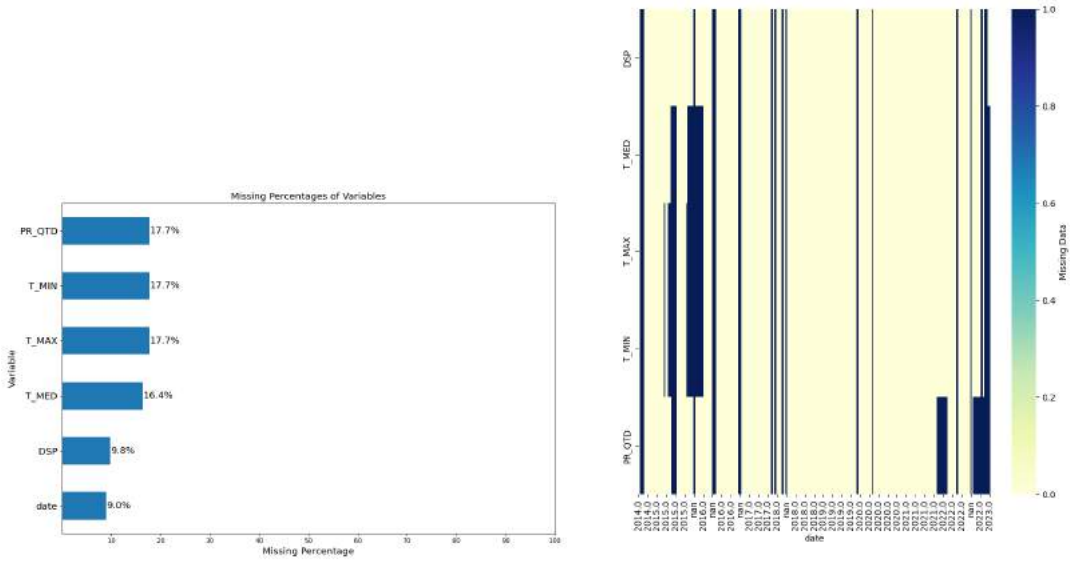
Figure A.17: Occurrences in each assigned sampling zone in the L7c2 region.



(a) Clean location distribution.

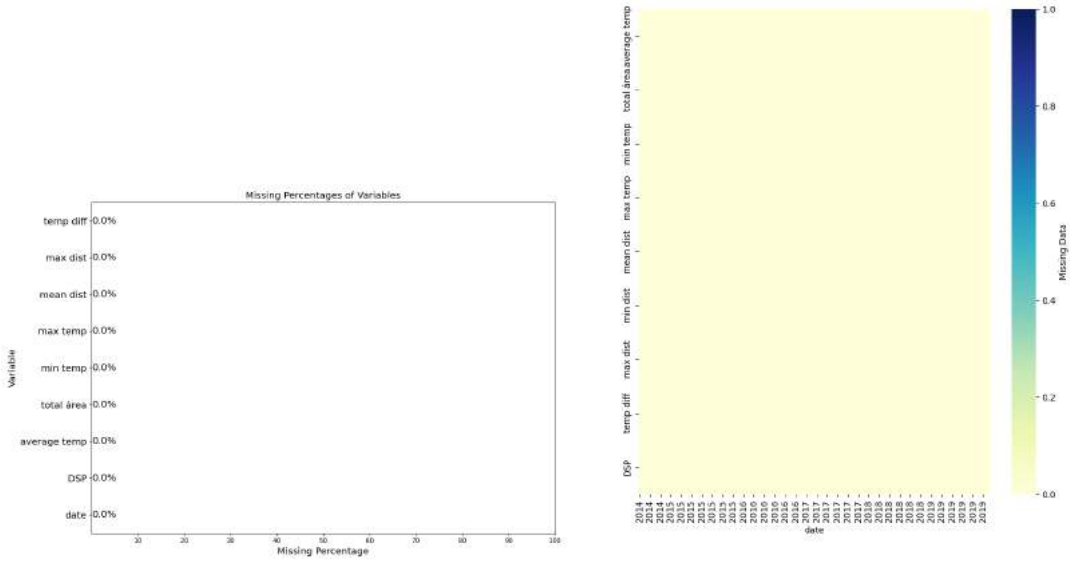
(b) Species distribution.

Figure A.18: L7c2 Porto de Mós clean location and species distributions.



(a) Missing percentages. (b) Missing periods.

Figure A.19: L7c2 Porto de Mós meteorological missing values.



(a) Missing percentages. (b) Missing periods.

Figure A.20: L7c2 Porto de Mós upwelling missing values.

A.2 Data Analysis

A.2.1 L2 Leça da Palmeira

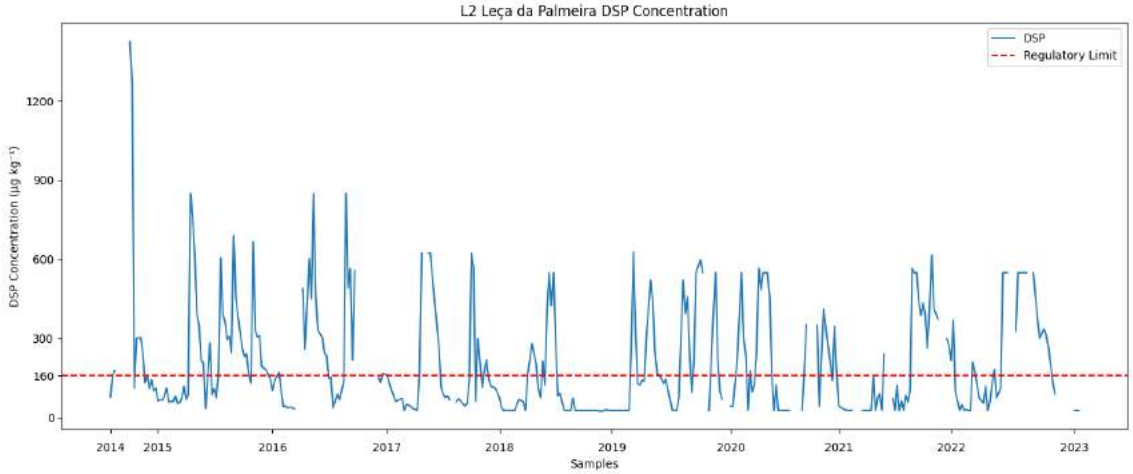
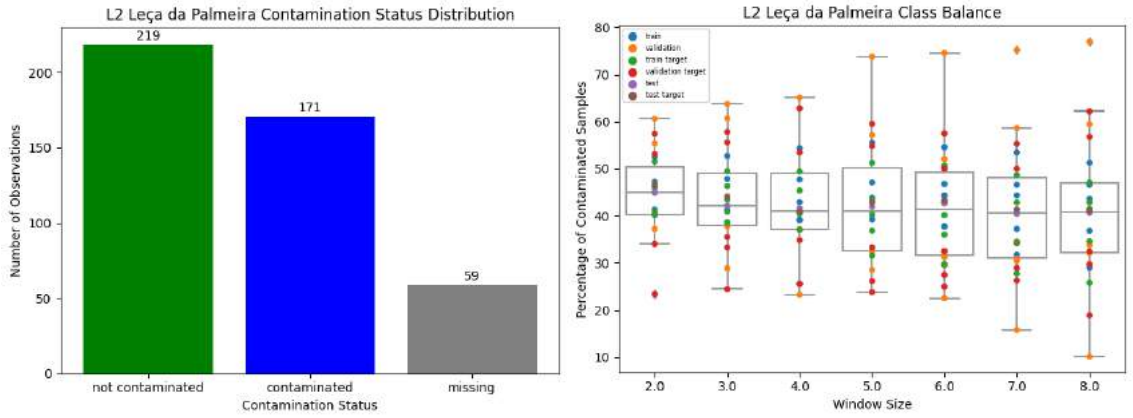


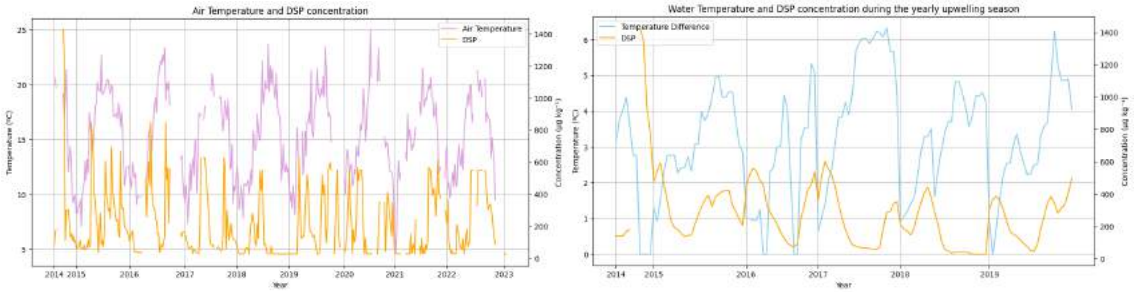
Figure A.21: L2 Leça da Palmeira DSP Concentration.



(a) Class distribution.

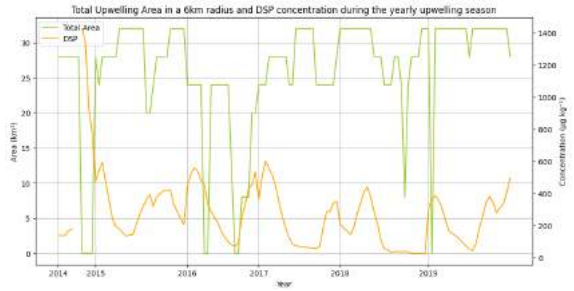
(b) Boxplot class balance.

Figure A.22: L2 Leça da Palmeira contamination class.



(a) Air Temperature.

(b) Temperature Difference.



(c) Total upwelling area.

Figure A.23: L2 Leça da Palmeira Environmental Variables.

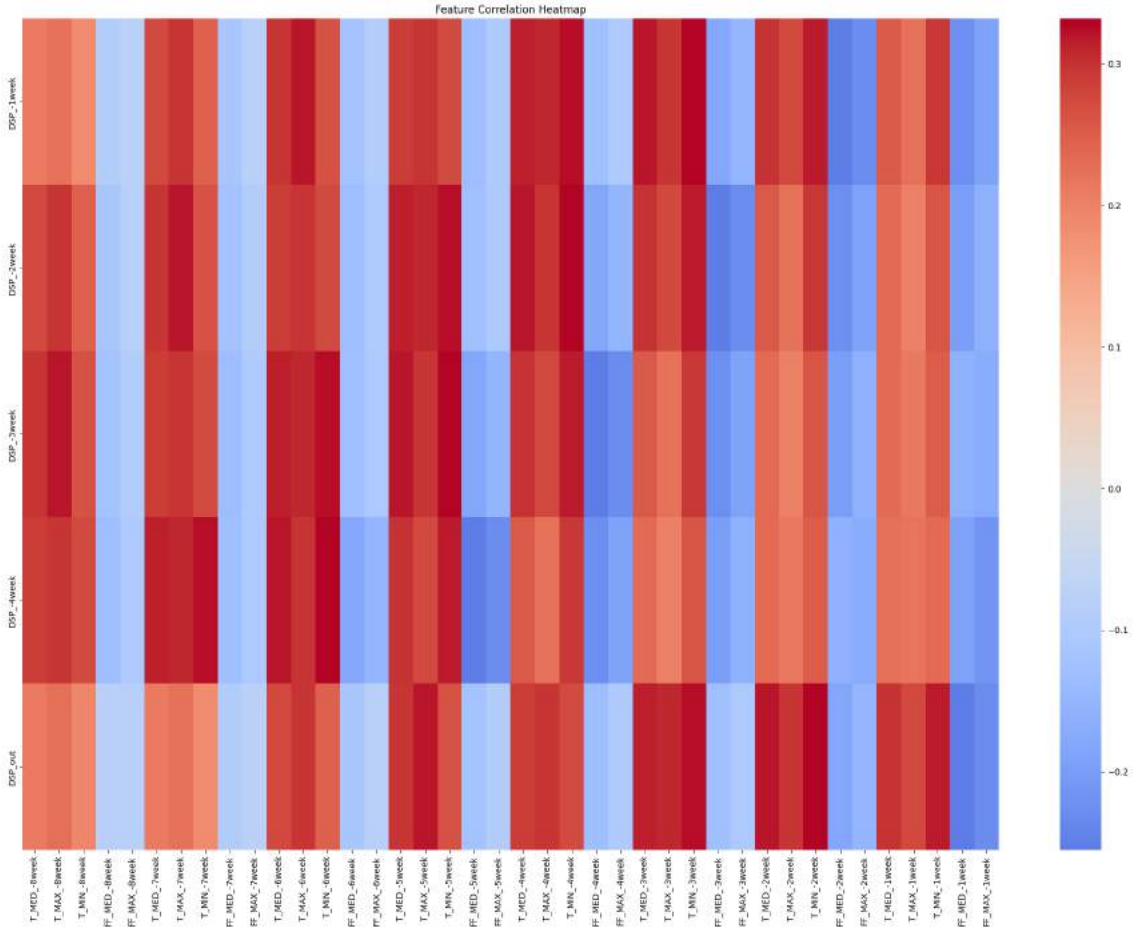


Figure A.24: L2 Leça da Palmeira Feature Correlation heatmap.

A.2.2 L5b Caparica

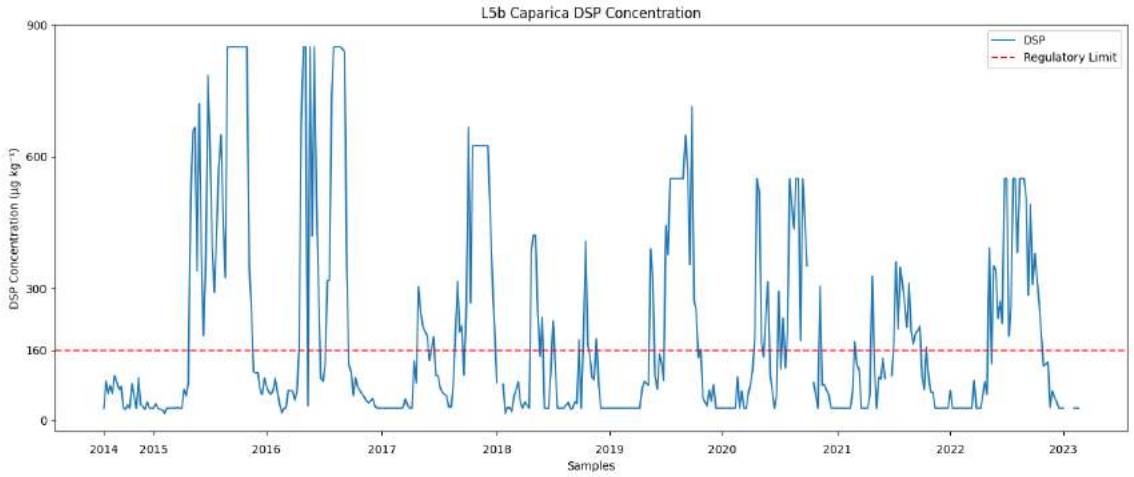
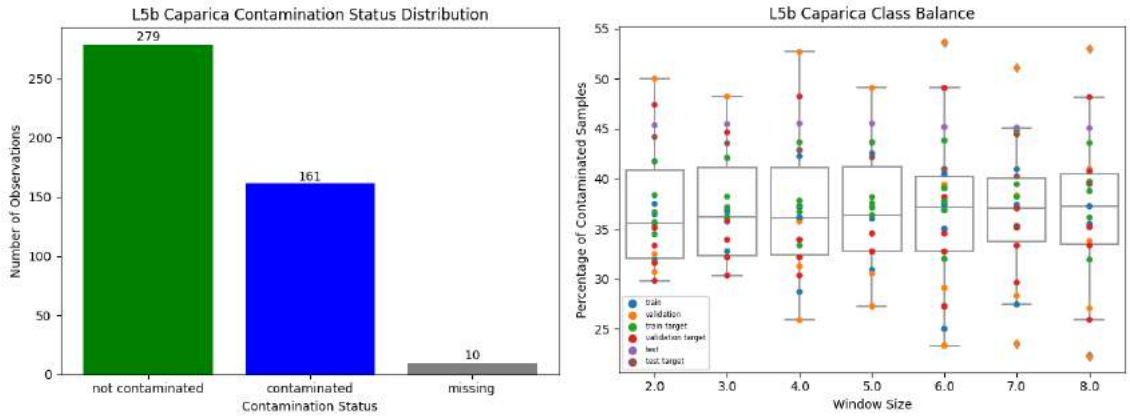


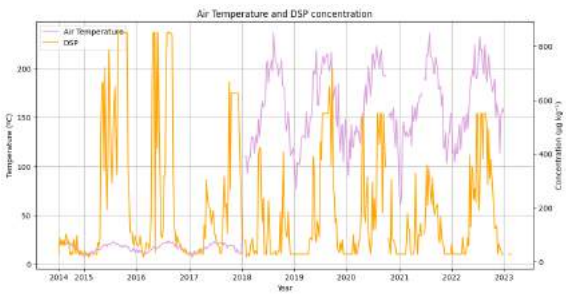
Figure A.25: L5b Caparica DSP Concentration.



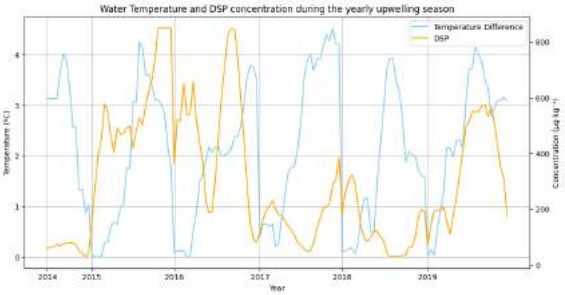
(a) Class distribution.

(b) Boxplot class balance.

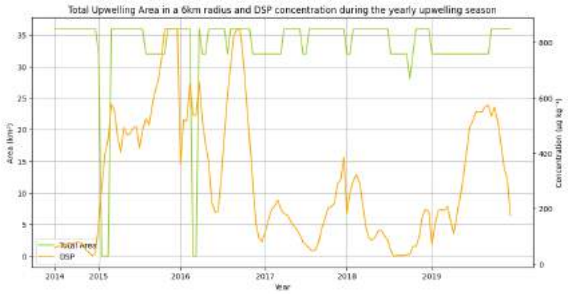
Figure A.26: L5b Caparica contamination class.



(a) Air Temperature.



(b) Temperature Difference.



(c) Total upwelling area.

Figure A.27: L5b Caparica Environmental Variables.

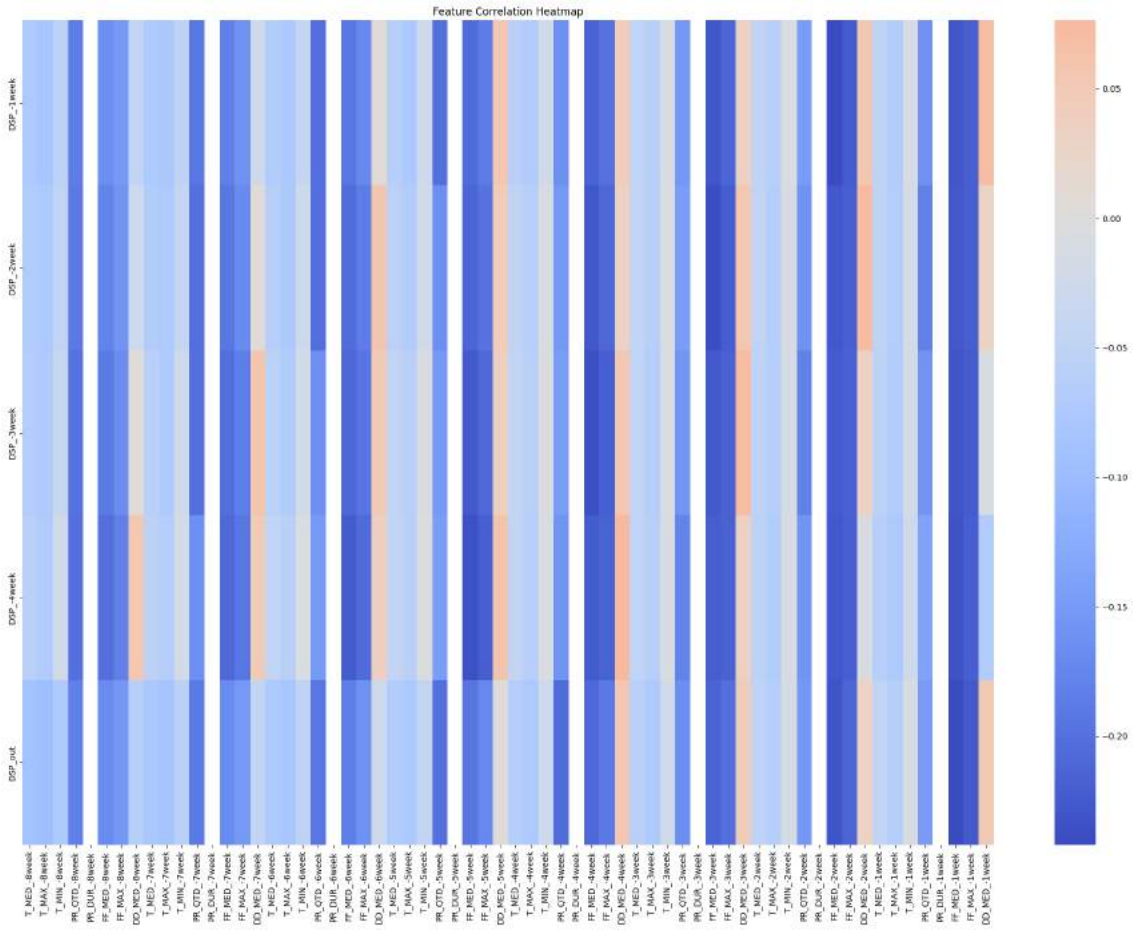


Figure A.28: L5b Caparica Feature Correlation heatmap.

A.2.3 RIAV1 Triângulo

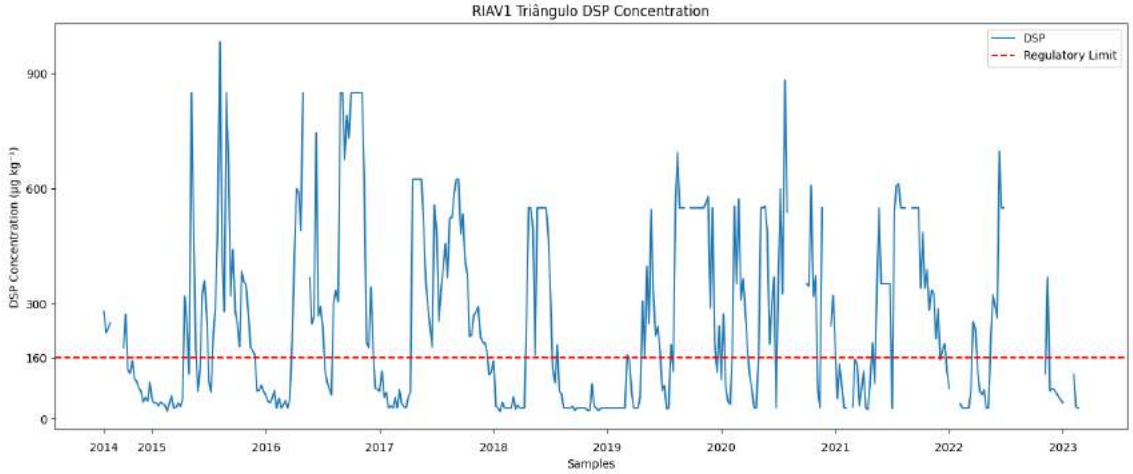
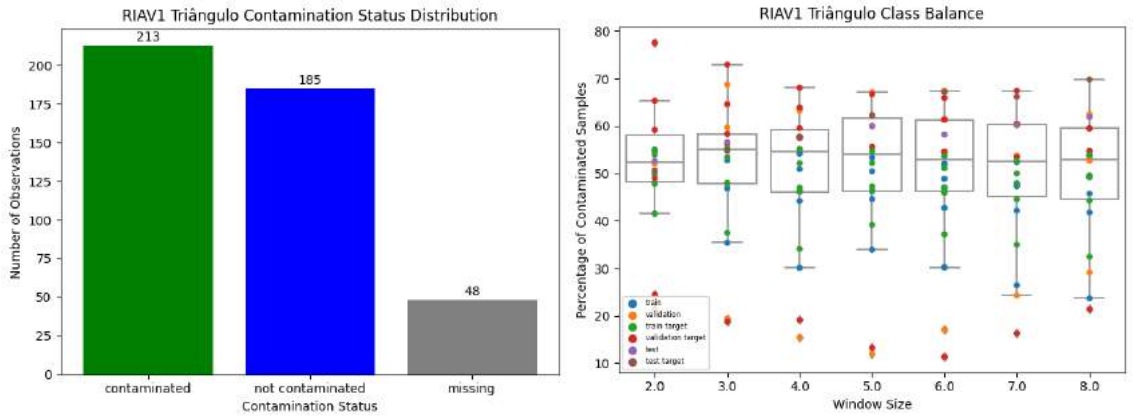


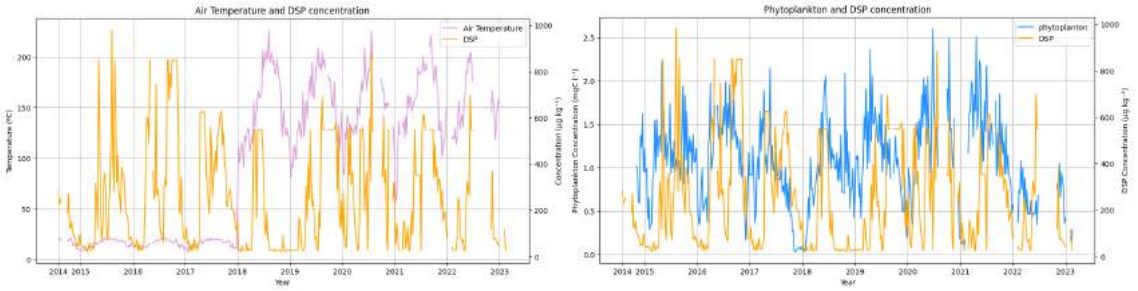
Figure A.29: RIAV1 Triângulo DSP Concentration.



(a) Class distribution.

(b) Boxplot class balance.

Figure A.30: RIAV1 Triângulo contamination class.



(a) Air Temperature. (b) Phytoplankton.

Figure A.31: RIAV1 Triângulo Environmental Variables.

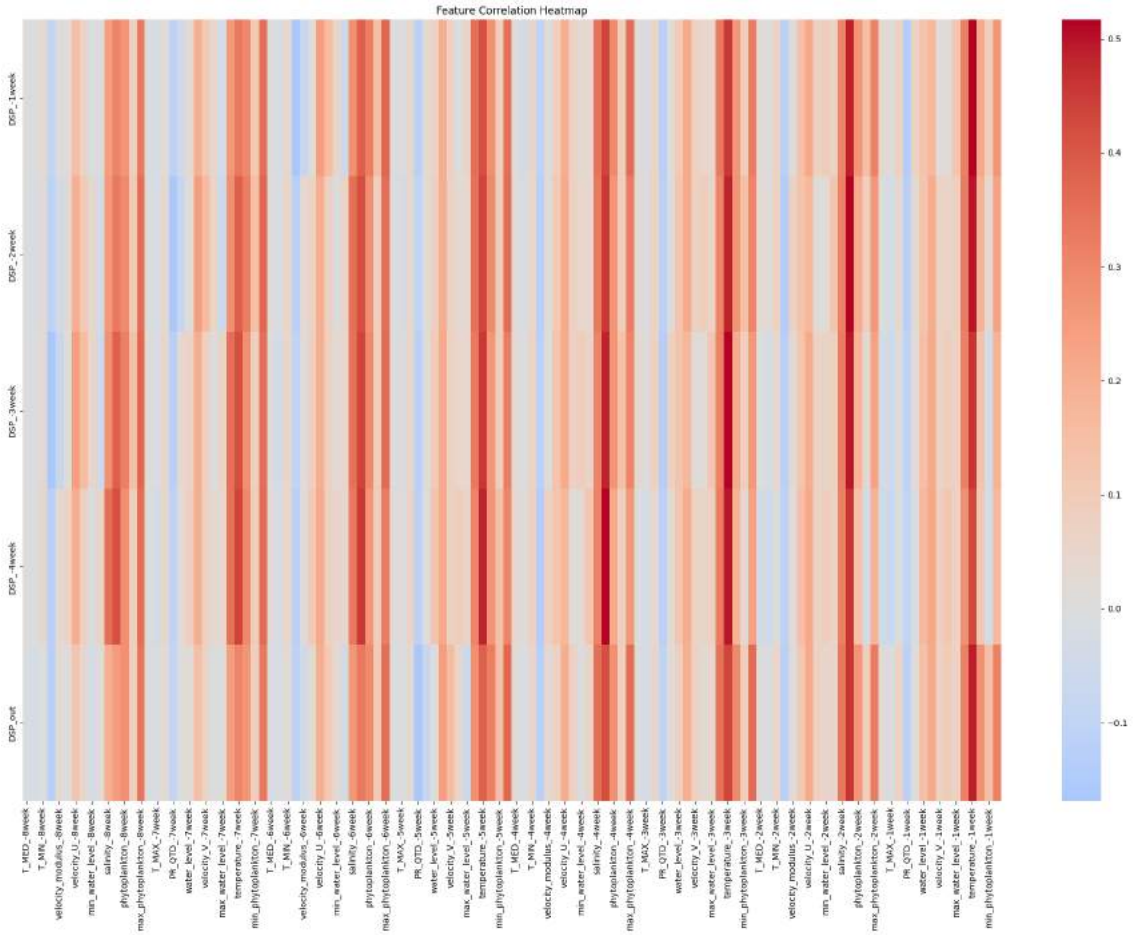


Figure A.32: RIAV1 Triângulo Feature Correlation heatmap.

A.2.4 L7c2 Porto de Mós

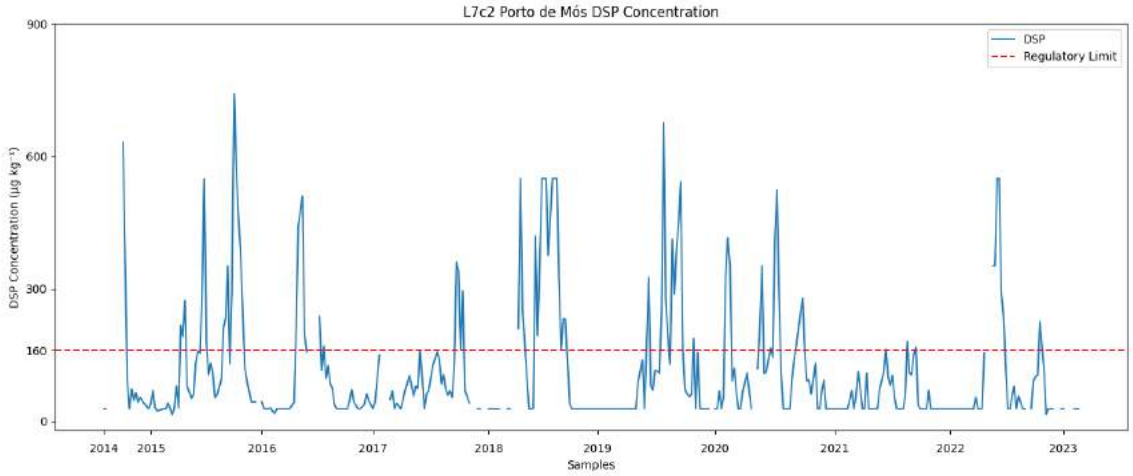
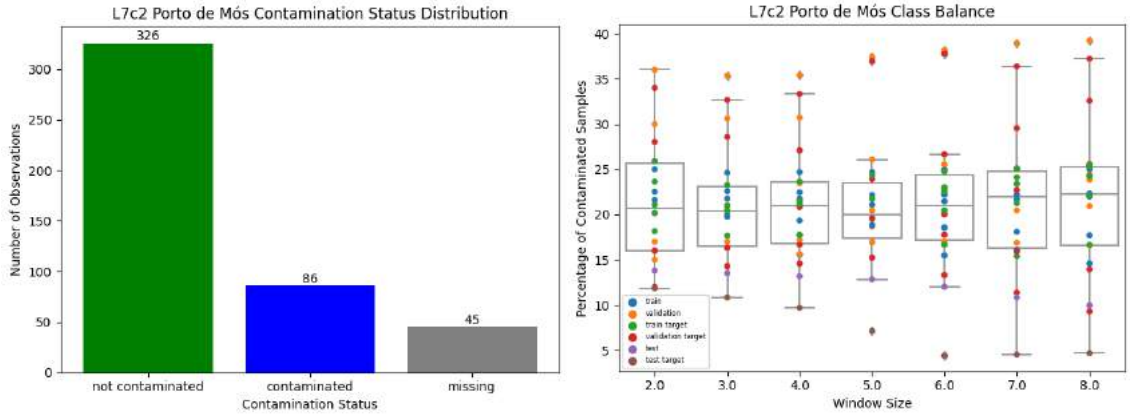


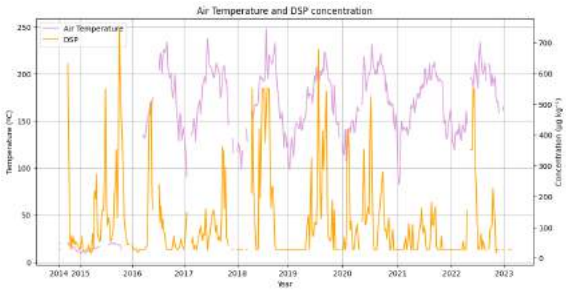
Figure A.33: L7c2 Porto de Mós DSP Concentration.



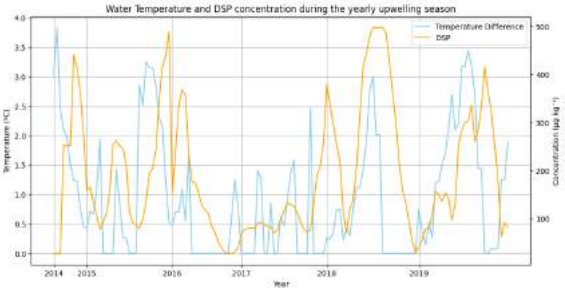
(a) Class distribution.

(b) Boxplot class balance.

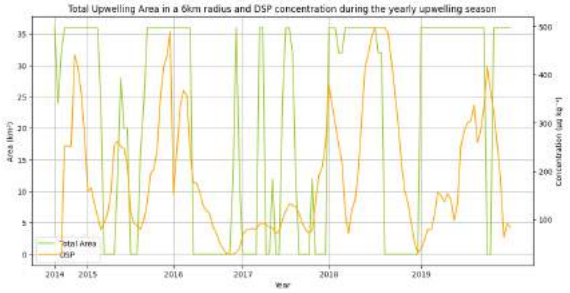
Figure A.34: L7c2 Porto de Mós contamination class.



(a) Air Temperature.



(b) Temperature Difference.



(c) Total upwelling area.

Figure A.35: L7c2 Porto de Mós Environmental Variables.

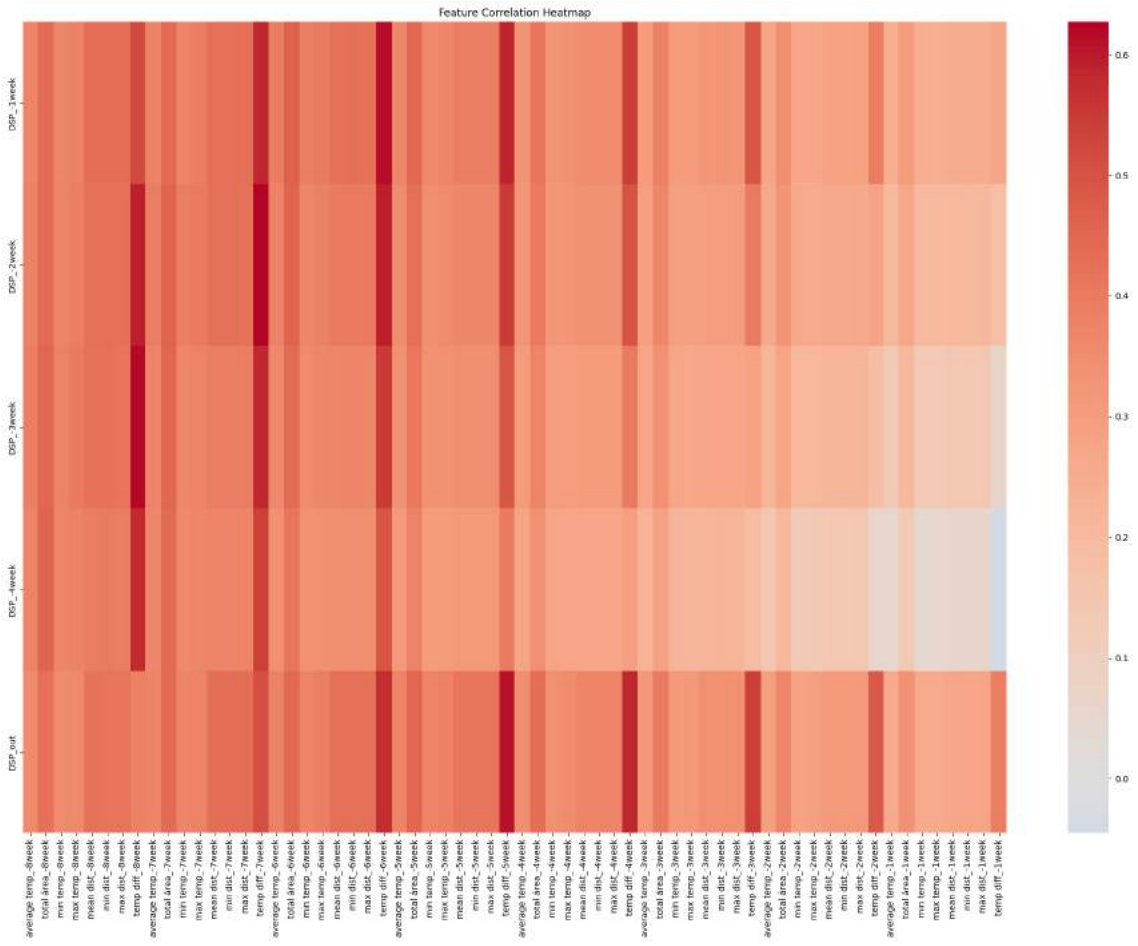


Figure A.36: L7c2 Porto de Mós Feature Correlation heatmap.

A.3 Additional Tables

Table A.1: Datasets Designation

Sampling Region	Dataset Designation	Upwelling Format	Included Variables			
			DSP	Meteorological	HWP	Upwelling
L1 Carreço	L1-D	-	✓	-	-	-
	L1-DM	-	✓	✓	-	-
	L1-DH	-	✓	-	✓	-
	L1-DMH	-	✓	✓	✓	-
	L1-UP-D	✓	✓	-	-	-
	L1-UP-DU	✓	✓	-	-	✓
	L1-UP-DM	✓	✓	✓	-	-
	L1-UP-DMU	✓	✓	✓	-	✓
	L1-UP-DH	✓	✓	-	✓	-
	L1-UP-DHU	✓	✓	-	✓	✓
	L1-UP-DMH	✓	✓	✓	✓	-
	L1-UP-DMHU	✓	✓	✓	✓	✓
L2 Leça da Palmeira	L2-D	-	✓	-	-	-
	L2-DM	-	✓	✓	-	-
	L2-UP-D	✓	✓	-	-	-
	L2-UP-DU	✓	✓	-	-	✓
	L2-UP-DM	✓	✓	✓	-	-
	L2-UP-DMU	✓	✓	✓	-	✓
L5b Caparica	L5b-D	-	✓	-	-	-
	L5b-DM	-	✓	✓	-	-
	L5b-UP-D	✓	✓	-	-	-
	L5b-UP-DU	✓	✓	-	-	✓
	L5b-UP-DM	✓	✓	✓	-	-
	L5b-UP-DMU	✓	✓	✓	-	✓
RIAV1 Triângulo	RIAV1-D	-	✓	-	-	-
	RIAV1-DM	-	✓	✓	-	-
	RIAV1-DH	-	✓	-	✓	-
	RIAV1-DMH	-	✓	✓	✓	-
L7c2 Porto de Mós	L7c2-UP-D	✓	✓	-	-	-
	L7c2-UP-DU	✓	✓	-	-	✓
	L7c2-UP-DM	✓	✓	✓	-	-
	L7c2-UP-DMU	✓	✓	✓	-	✓

Table A.2: Meteorological Variables

Code	Name	Unit
P_E_MD	Average atmospheric pressure at station level	hPa
P_E_MX	Maximum atmospheric pressure at station level	hPa
P_E_MN	Minimum atmospheric pressure at station level	hPa
P_E_DP	Standard deviation of atmospheric pressure at station level	hPa
P_M_MD	Average atmospheric pressure at the mean sea level	hPa
P_M_MX	Maximum atmospheric pressure at the mean sea level	hPa
P_M_MN	Minimum atmospheric pressure at the mean sea level	hPa
P_M_DP	Standard deviation of atmospheric pressure at the mean sea level	hPa
PR_V3h	Atmospheric pressure variation in the previous 3 hours	hPa
PR_TDC	Atmospheric pressure trend in the previous 3 hours	Numerical
T_INST	Instantaneous air temperature at 1.5 meters	ºC
T_MAX	Maximum air temperature at 1.5 meters	ºC
T_MIN	Minimum air temperature at 1.5 meters	ºC
T_MED	Mean air temperature at 1.5 meters	ºC
T_DP	Standard deviation of air temperature at 1.5 meters	ºC
HR_INS	Instantaneous relative humidity	%
HR_MED	Mean relative humidity	%
HR_MAX	Maximum relative humidity	%
HR_MIN	Minimum relative humidity	%
HR_DP	Standard deviation of relative humidity	%
TW_INS	Instantaneous wet-bulb temperature	ºC
TW_MED	Mean wet-bulb temperature	ºC
TW_MAX	Maximum wet-bulb temperature	ºC
TW_MIN	Minimum wet-bulb temperature	ºC
TW_DP	Standard deviation of the wet-bulb temperature	ºC
TD_INS	Instantaneous dew point temperature	ºC
TD_MED	Mean dew point temperature	ºC
TD_MAX	Maximum dew point temperature	ºC
TD_MIN	Minimum dew point temperature	m
TD_DP	Standard deviation of the dew point temperature	ºC
E_INS	Instantaneous vapor pressure	hPa
E_MED	Mean vapor pressure	hPa
E_MAX	Maximum vapor pressure	hPa
E_MIN	Minimum vapor pressure	hPa
E_DP	Standard deviation of the vapor pressure	hPa
DD_FFX	Maximum wind direction	º
DD_DP	Standard deviation of wind direction	º
DD_MED	Mean wind direction	º
DD_PRE	Predominant wind direction	Categorical
FF_MAX	Maximum wind intensity	m/s
FF_MED	Mean wind intensity	m/s
FF_DP	Standard deviation of the wind intensity	m/s
TSUP_M	Mean air temperature at +0.05 meters	ºC
TSUP_X	Maximum air temperature at +0.05 meters	ºC
TSUP_N	Minimum air temperature at +0.05 meters	ºC
TSUP_D	Standard deviation of the air temperature at +0.05 meters	ºC
T_05_M	Mean air temperature at -0.05 meters	ºC
T_05_X	Maximum air temperature at -0.05 meters	ºC
T_05_N	Minimum air temperature at -0.05 meters	ºC
T_05_D	Standard deviation of the air temperature at -0.05 meters	ºC
T_10_M	Mean air temperature at -0.10 meters	ºC
T_10_X	Maximum air temperature at -0.10 meters	ºC
T_10_N	Minimum air temperature at -0.10 meters	ºC
T_10_D	Standard deviation of the air temperature at -0.10 meters	ºC
T_20_M	Mean air temperature at -0.20 meters	ºC
T_20_X	Maximum air temperature at -0.20 meters	ºC
T_20_N	Minimum air temperature at -0.20 meters	ºC
T_20_D	Standard deviation of the air temperature at -0.20 meters	ºC
T_50_M	Mean air temperature at -0.50 meters	ºC
T_100_M	Mean air temperature at -1.0 meters	ºC
PR_DUR	Precipitation duration	Minutes
PR_QTD	Precipitation amount	mm
PR_I_X	Maximum intensity of instantaneous precipitation	mm/h
RG_TOT	Total global radiation	KJ/m^2
RG_MAX	Maximum global radiation in a minute	W/M^2
RG_MIN	Minimum global radiation in a minute	W/M^2
RD_TOT	Total diffuse radiation	KJ/m^2
RD_MAX	Maximum diffuse radiation in a minute	W/M^2
RD_MIN	Minimum diffuse radiation in a minute	W/M^2

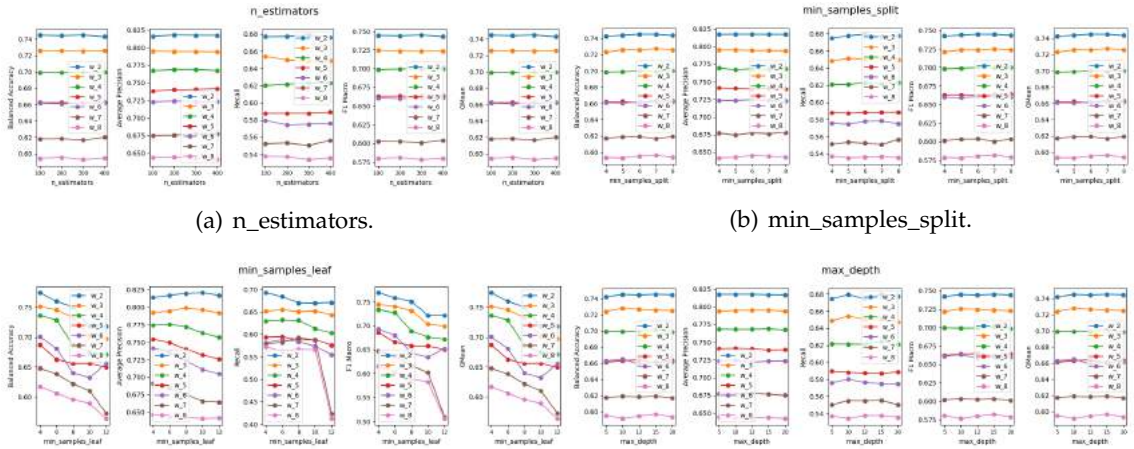
B

Appendix 2: Models Tuning

B.1 Classification

B.1.1 Random Forest

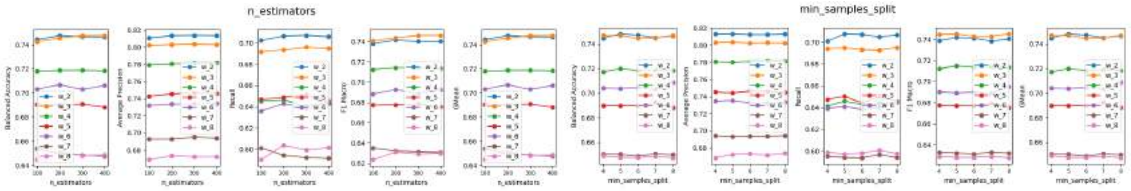
B.1.1.1 L1 Carreço RF Classification



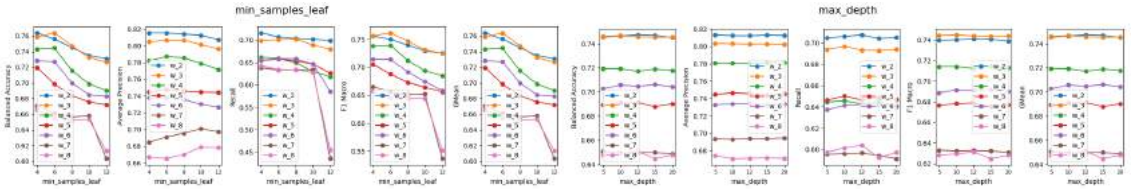
(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.1: RF hyperparameters tuning for the L1-DM model.

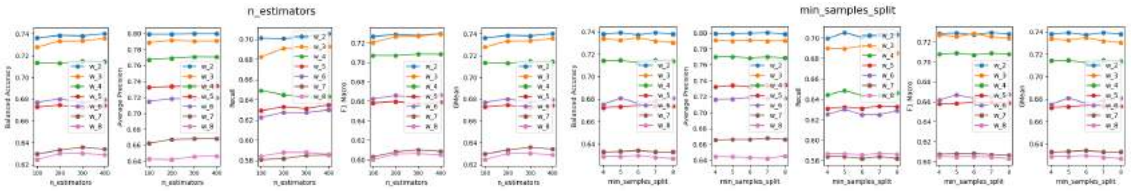


(a) n_estimators. (b) min_samples_split.

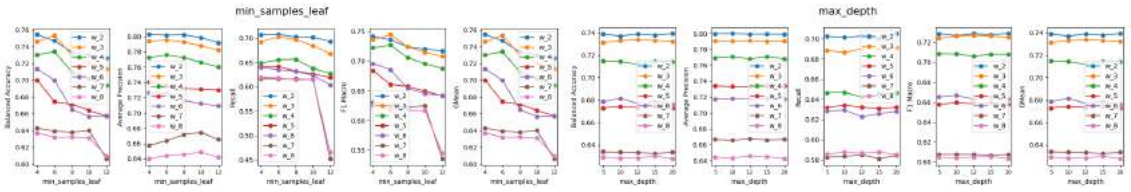


(c) min_samples_leaf. (d) max_depth.

Figure B.2: RF hyperparameters tuning for the L1-DH model.



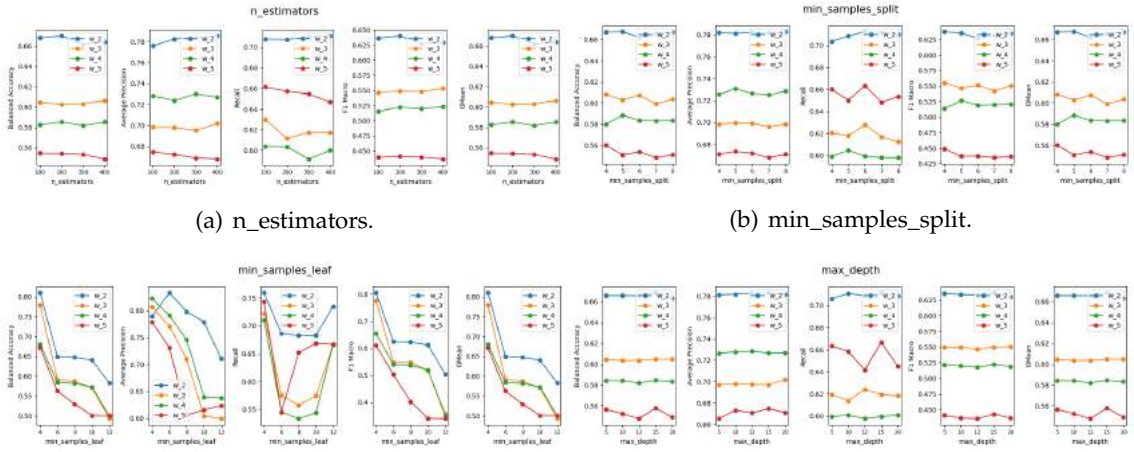
(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure B.3: RF hyperparameters tuning for the L1-DMH model.

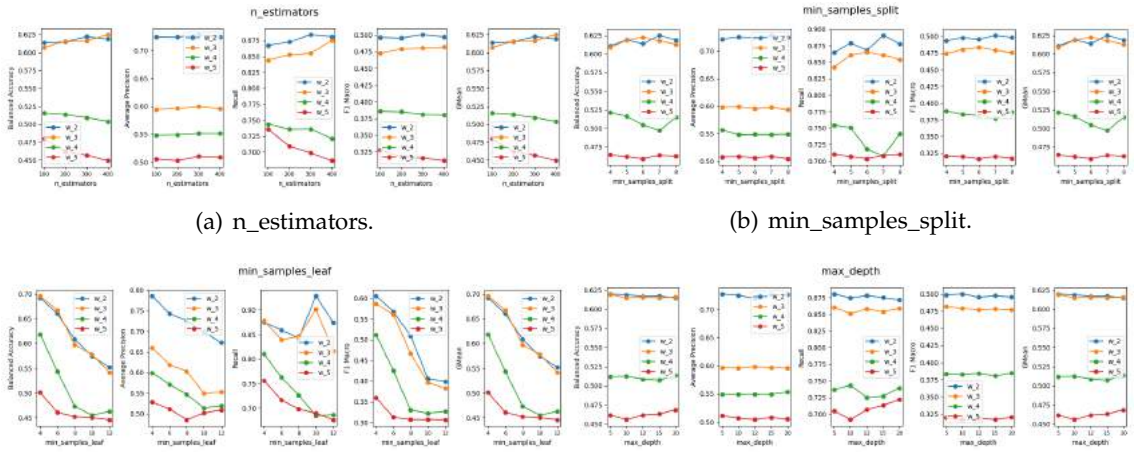
B.1.1.2 L1 Carreço RF Upwelling Classification



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

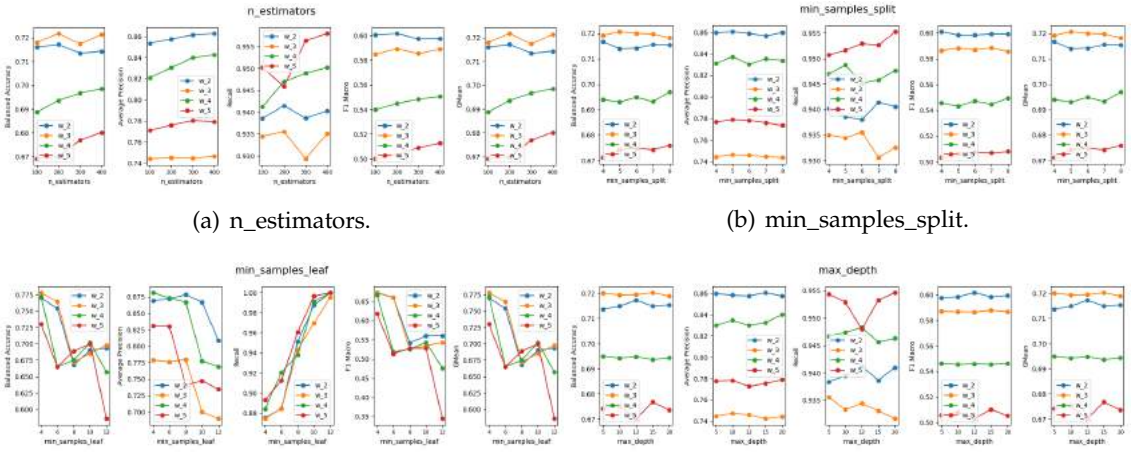
Figure B.4: RF hyperparameters tuning for the L1-UP-D model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

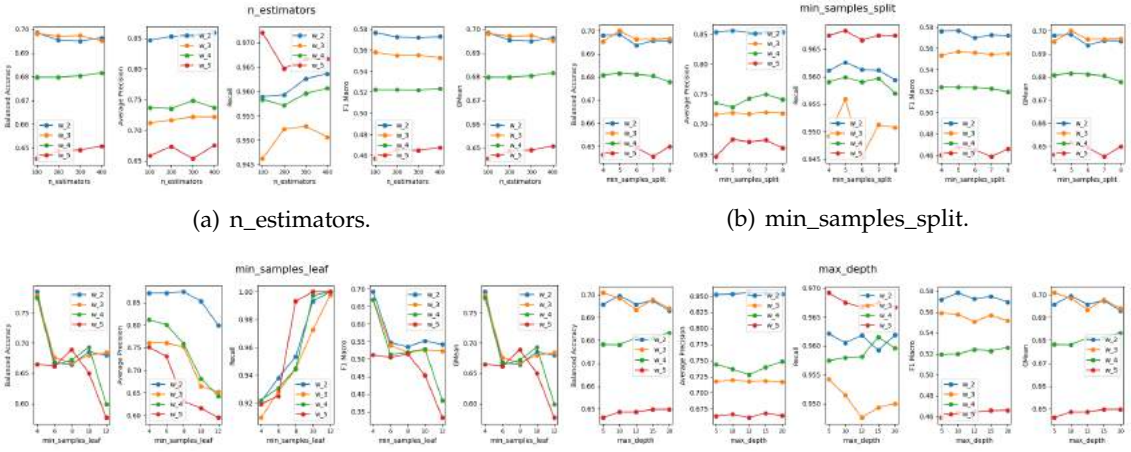
Figure B.5: RF hyperparameters tuning for the L1-UP-DU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

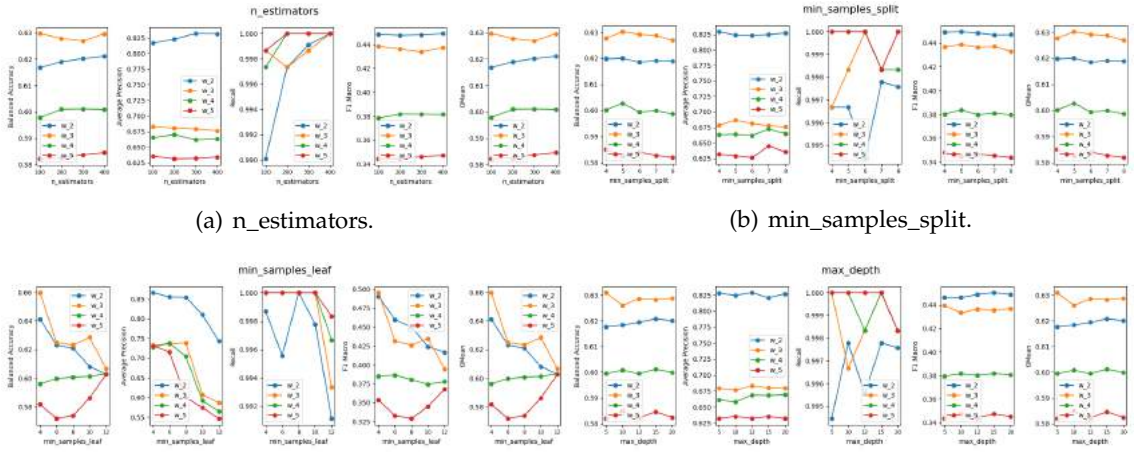
Figure B.6: RF hyperparameters tuning for the L1-UP-DM model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

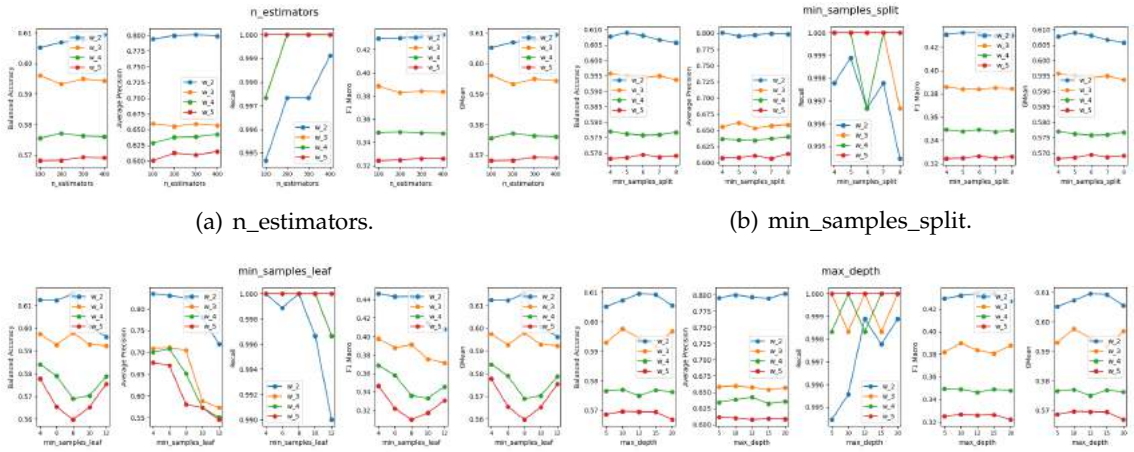
Figure B.7: RF hyperparameters tuning for the L1-UP-DMU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

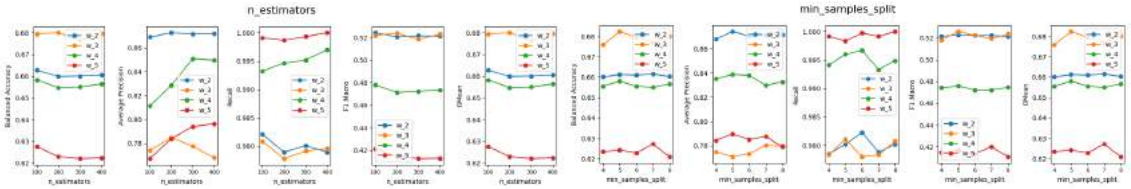
Figure B.8: RF hyperparameters tuning for the L1-UP-DH model.



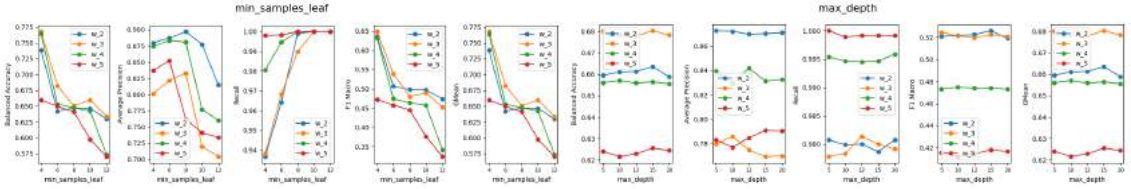
(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.9: RF hyperparameters tuning for the L1-UP-DHU model.

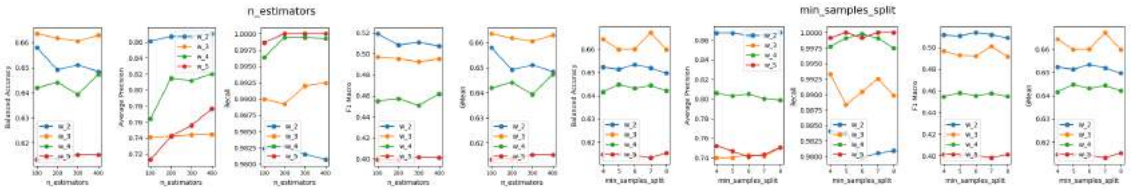


(a) n_estimators. (b) min_samples_split.

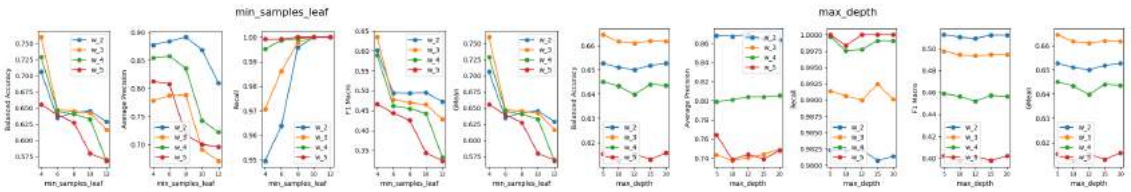


(c) min_samples_leaf. (d) max_depth.

Figure B.10: RF hyperparameters tuning for the L1-UP-DMH model.



(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure B.11: RF hyperparameters tuning for the L1-UP-DMHU model.

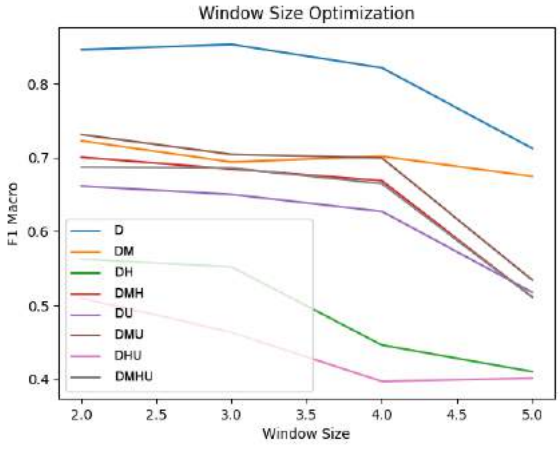
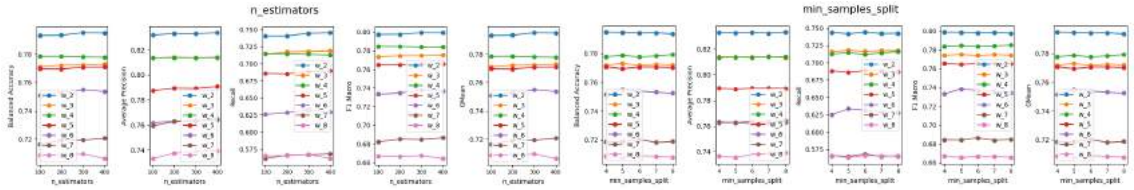
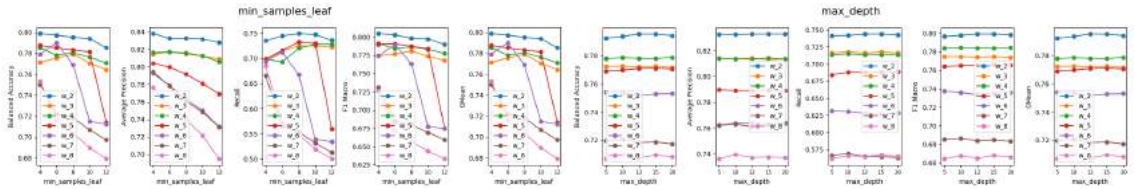


Figure B.12: L1 Carreço Window Optimization for RF Upwelling Classification models.

B.1.1.3 L2 Leça da Palmeira RF Classification

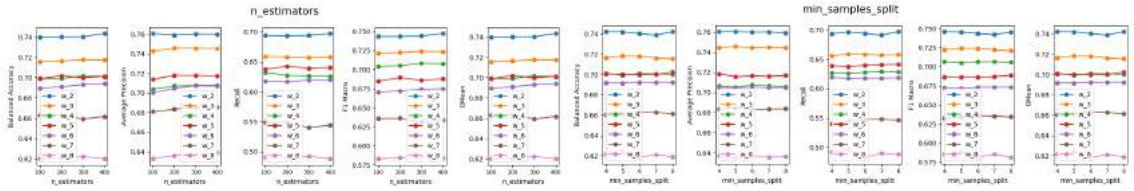


(a) n_estimators. (b) min_samples_split.

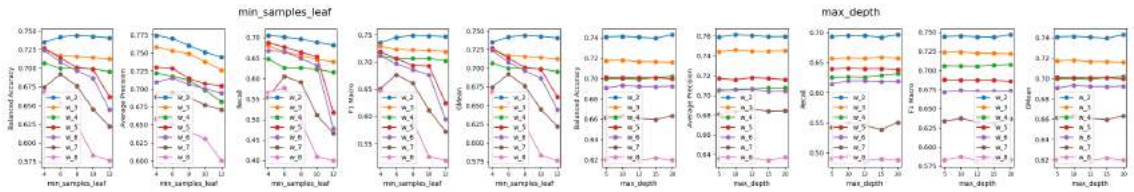


(c) min_samples_leaf. (d) max_depth.

Figure B.13: RF hyperparameters tuning for the L2-D model.



(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure B.14: RF hyperparameters tuning for the L2-DM model.

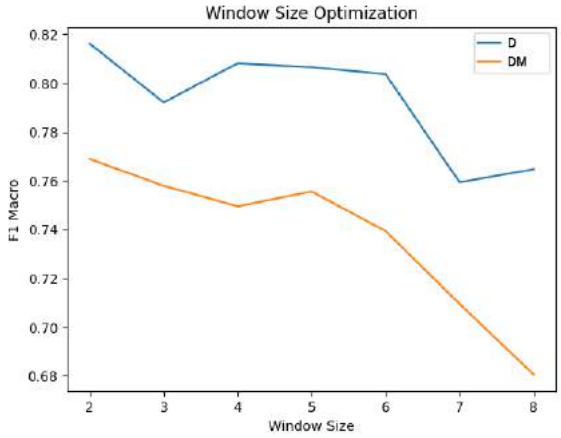
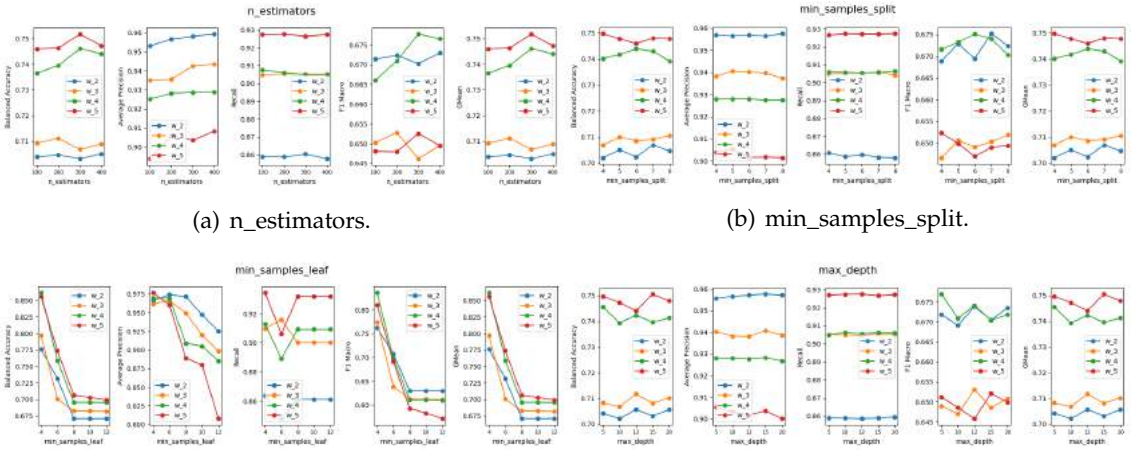


Figure B.15: L2 Leça da Palmeira Window Optimization for RF Classification models.

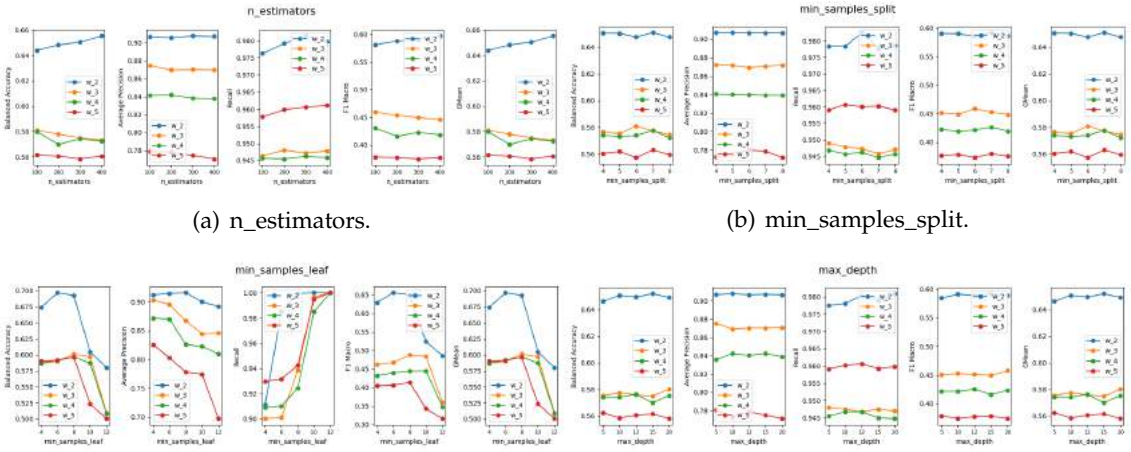
B.1.1.4 L2 Leça da Palmeira RF Upwelling Classification



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

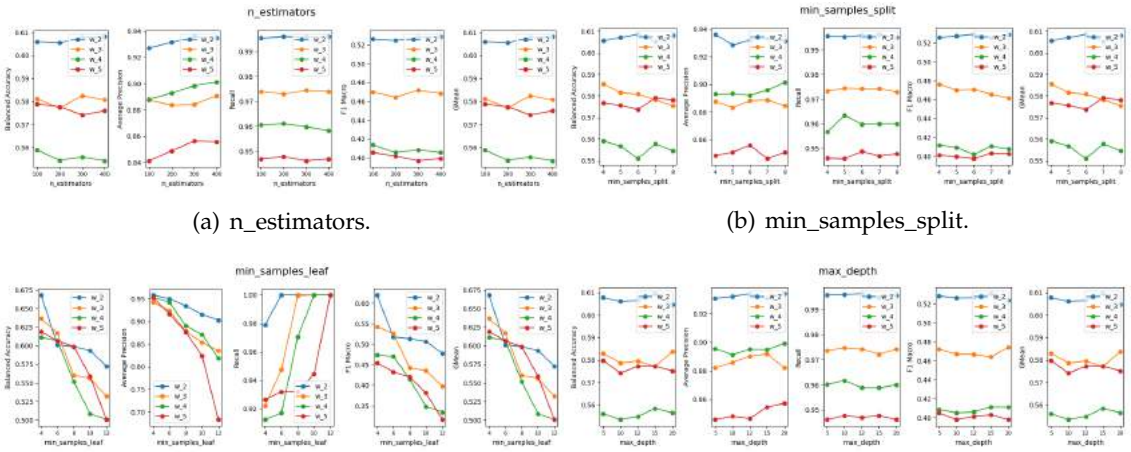
Figure B.16: RF hyperparameters tuning for the L2-UP-D model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

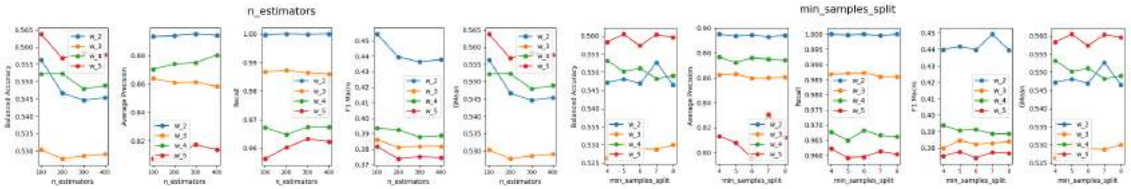
Figure B.17: RF hyperparameters tuning for the L2-UP-DU model.



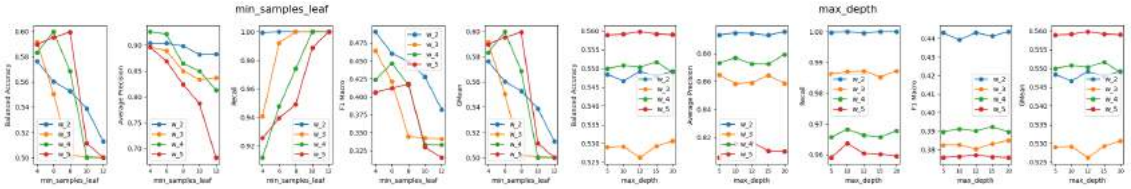
(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.18: RF hyperparameters tuning for the L2-UP-DM model.



(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure B.19: RF hyperparameters tuning for the L2-UP-DMU model.

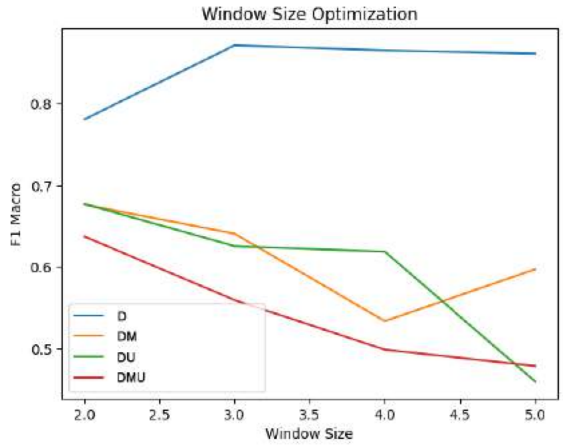
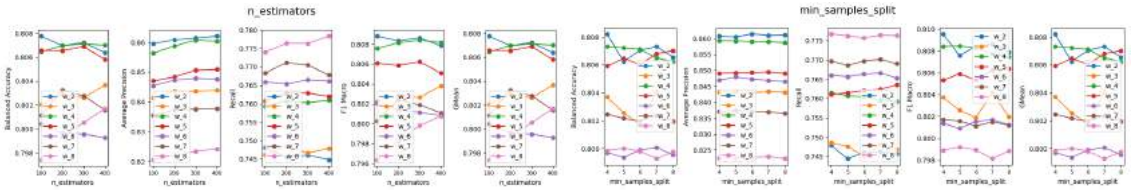
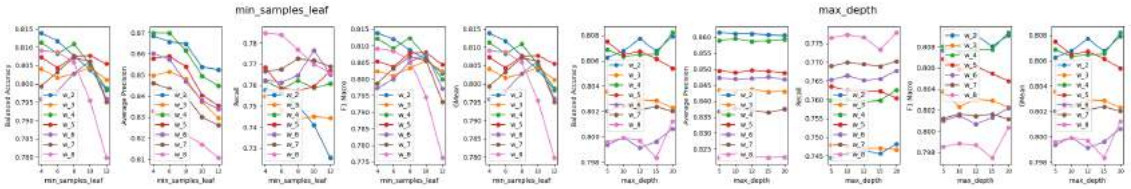


Figure B.20: L2 Leça da Palmeira Window Optimization for RF Upwelling Classification models.

B.1.1.5 L5b Caparica RF Classification

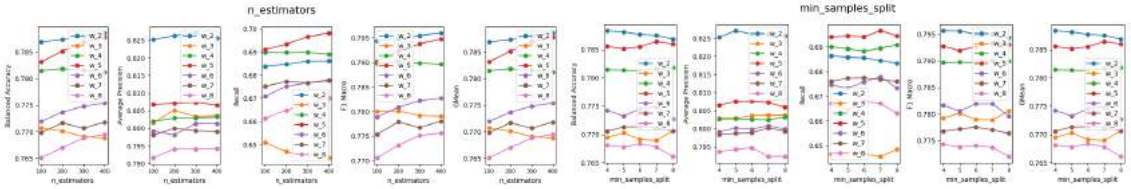


(a) n_estimators. (b) min_samples_split.

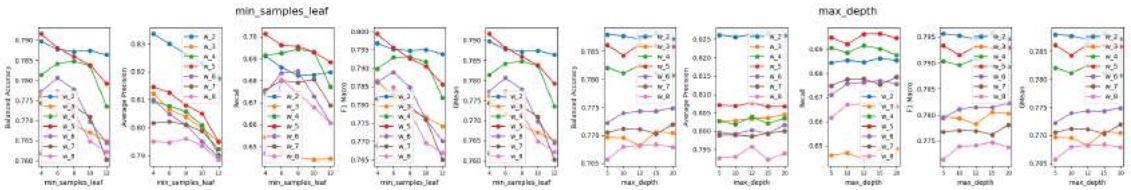


(c) min_samples_leaf. (d) max_depth.

Figure B.21: RF hyperparameters tuning for the L5b-D model.



(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure B.22: RF hyperparameters tuning for the L5b-DM model.

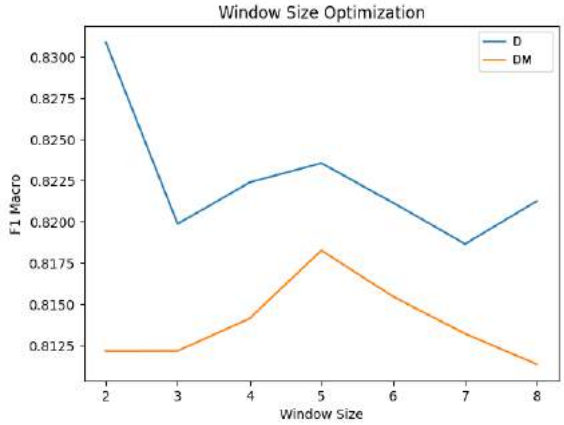
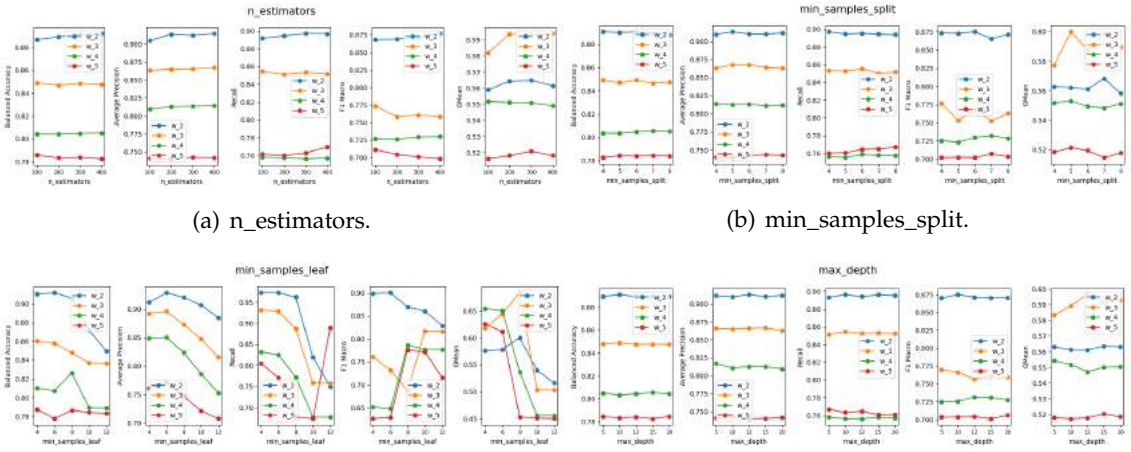


Figure B.23: L5b Caparica Window Optimization for RF Classification models.

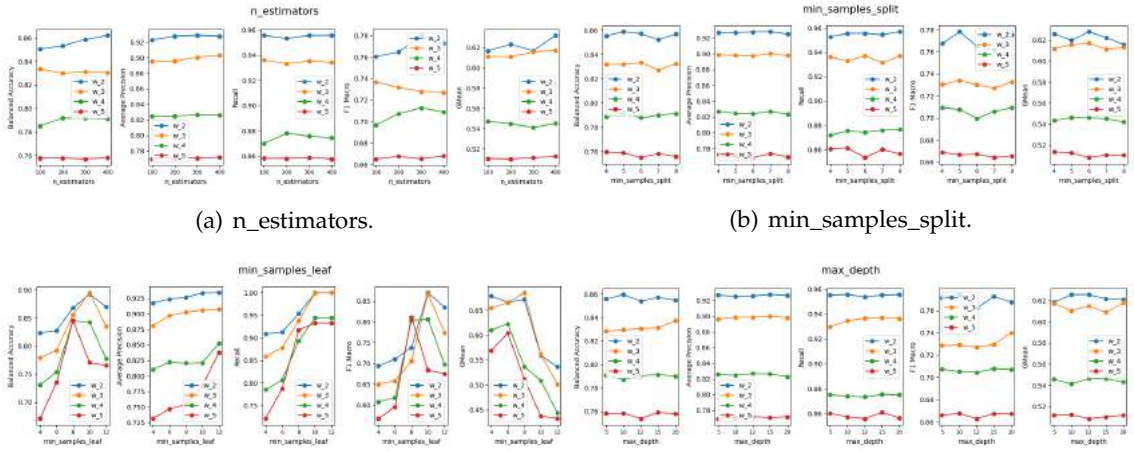
B.1.1.6 L5b Caparica RF Upwelling Classification



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

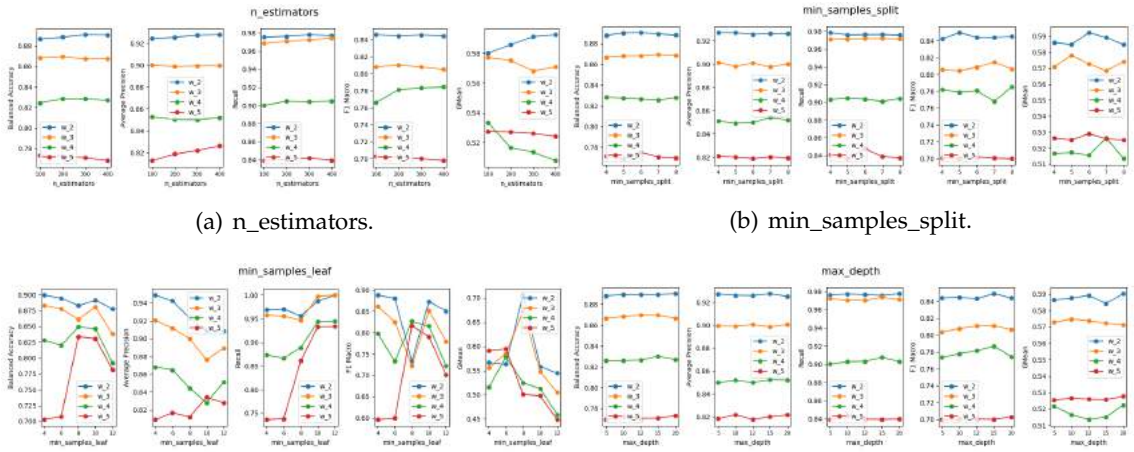
Figure B.24: RF hyperparameters tuning for the L5b-UP-D model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

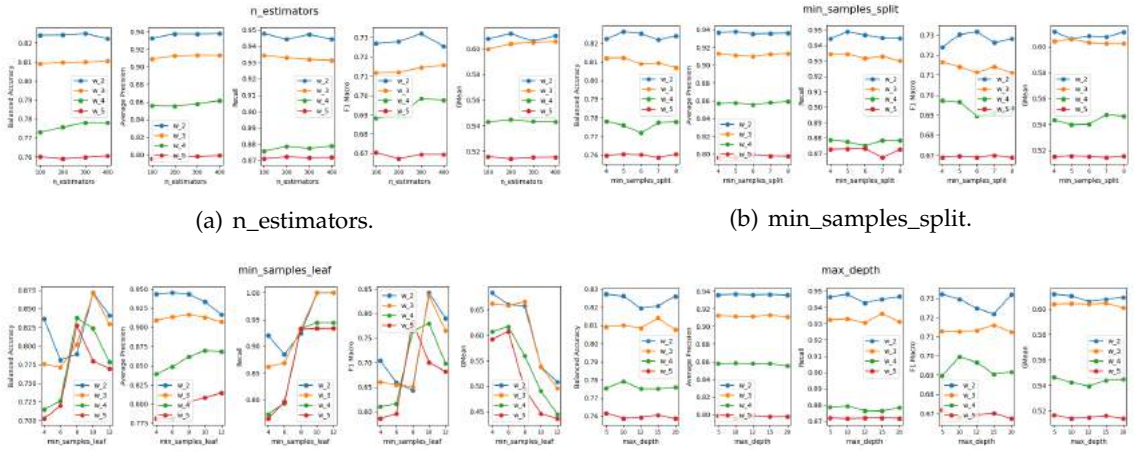
Figure B.25: RF hyperparameters tuning for the L5b-UP-DU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.26: RF hyperparameters tuning for the L5b-UP-DM model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.27: RF hyperparameters tuning for the L5b-UP-DMU model.

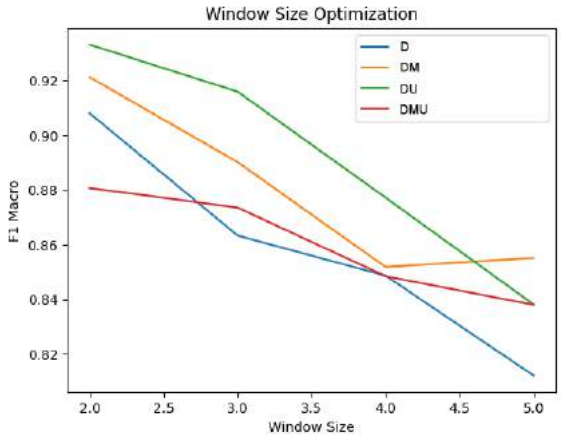
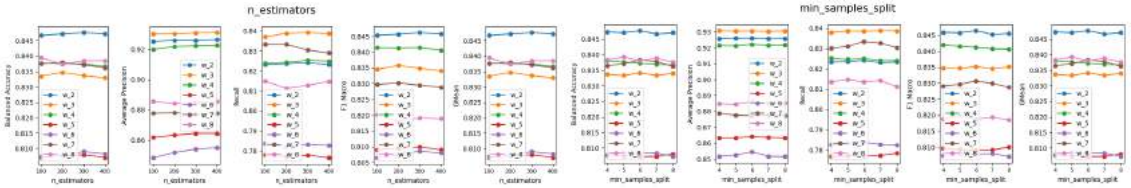
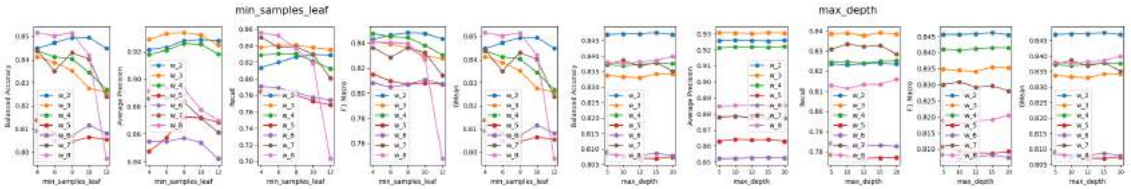


Figure B.28: L5b Caparica Window Optimization for RF Upwelling Classification models.

B.1.1.7 RIAV1 Triângulo RF Classification

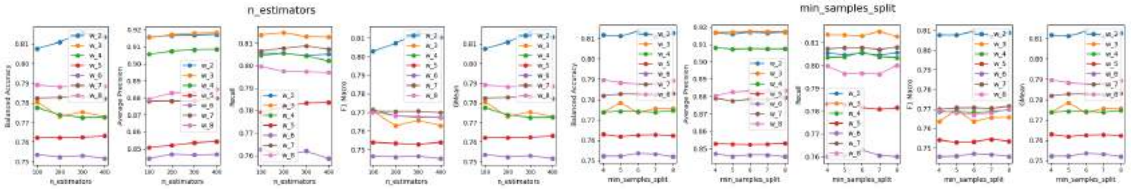


(a) n_estimators. (b) min_samples_split.

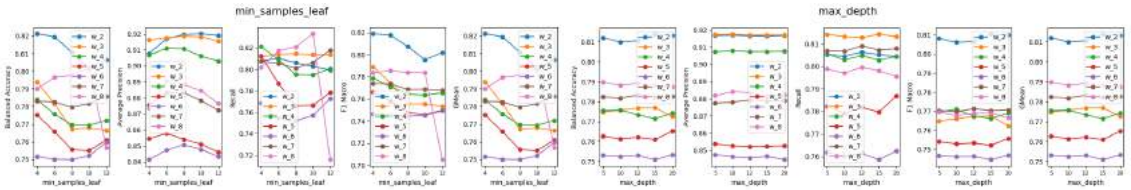


(c) min_samples_leaf. (d) max_depth.

Figure B.29: RF hyperparameters tuning for the RIAV1-D model.

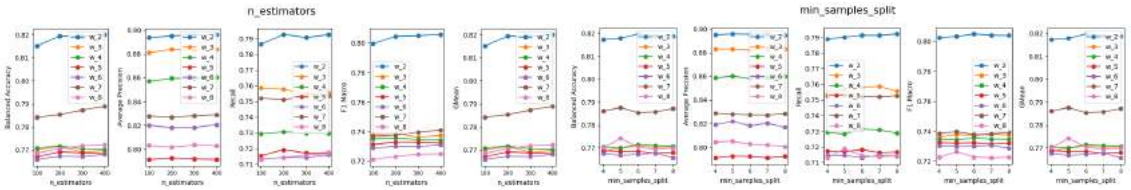


(a) n_estimators. (b) min_samples_split.

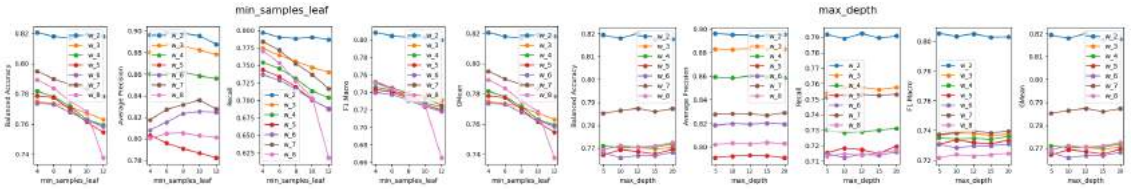


(c) min_samples_leaf. (d) max_depth.

Figure B.30: RF hyperparameters tuning for the RIAV1-DM model.

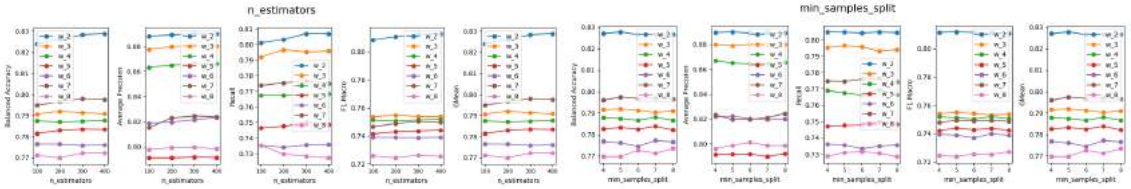


(a) n_estimators. (b) min_samples_split.

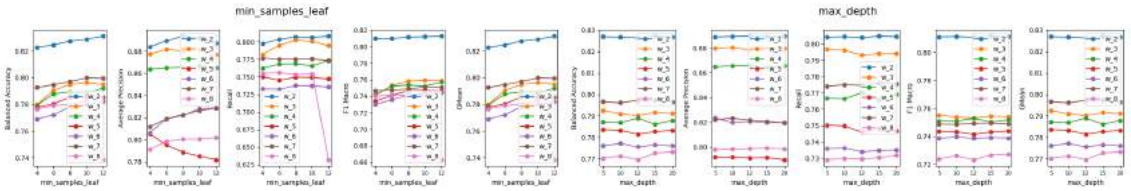


(c) min_samples_leaf. (d) max_depth.

Figure B.31: RF hyperparameters tuning for the RIAV1-DH model.



(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure B.32: RF hyperparameters tuning for the RIAV1-DMH model.

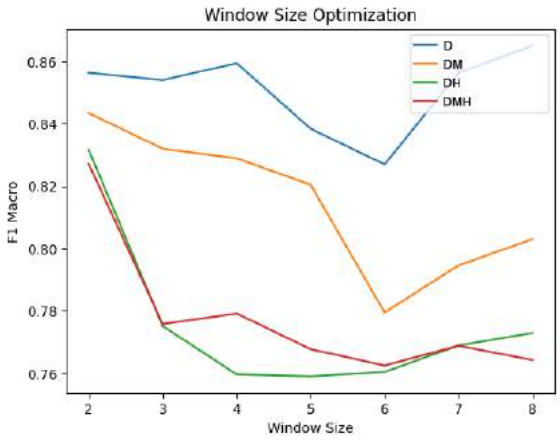
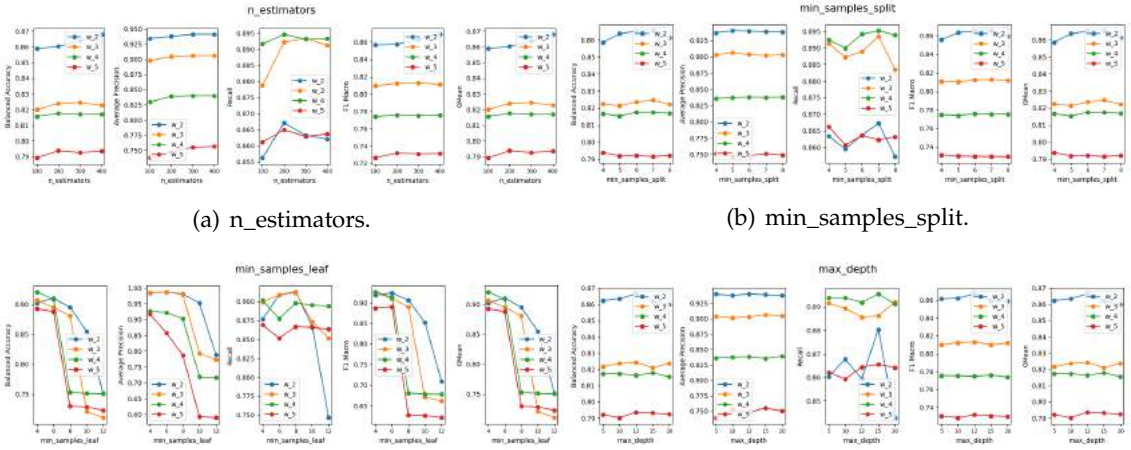


Figure B.33: RIAV1 Triângulo Window Optimization for RF Classification models.

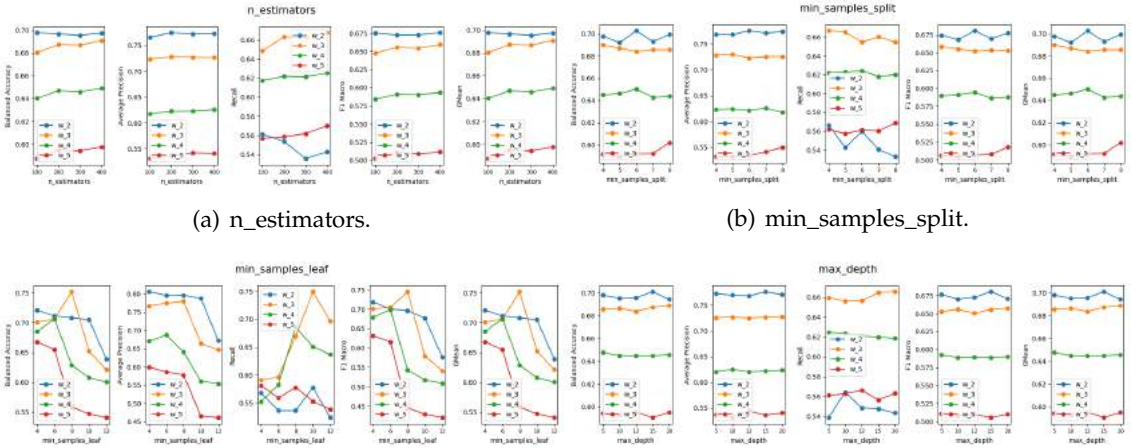
B.1.1.8 L7c2 Porto de Mós RF Upwelling Classification



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

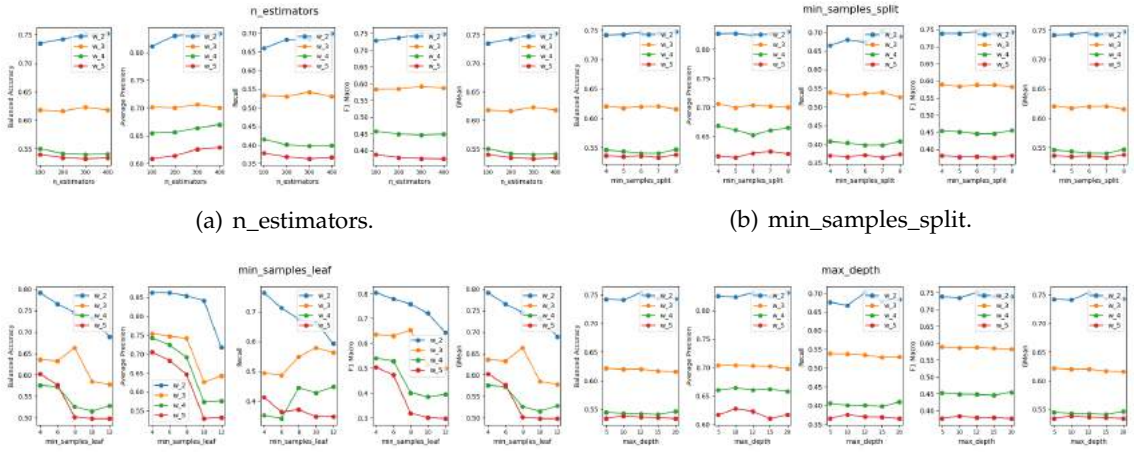
Figure B.34: RF hyperparameters tuning for the L7c2-UP-D model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

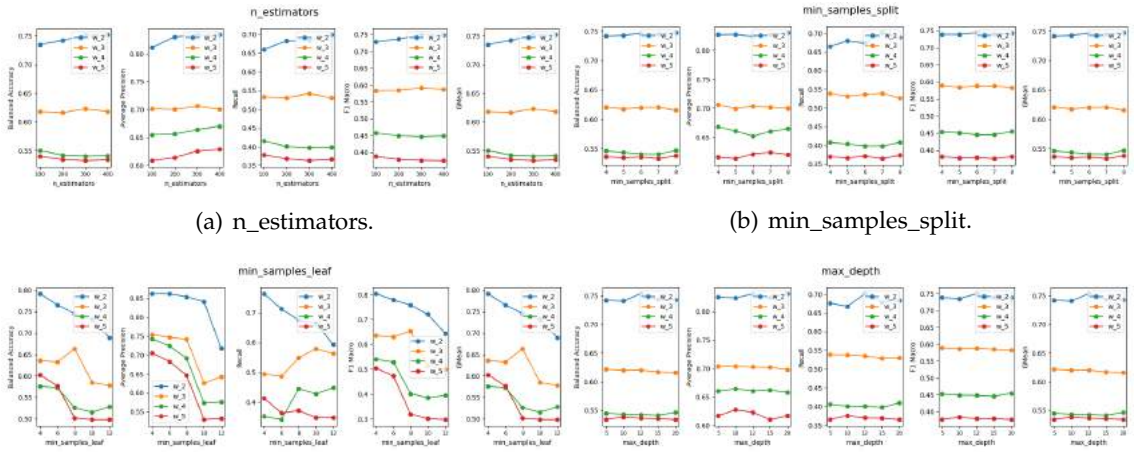
Figure B.35: RF hyperparameters tuning for the L7c2-UP-DU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.36: RF hyperparameters tuning for the L7c2-UP-DM model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.37: RF hyperparameters tuning for the L7c2-UP-DMU model.

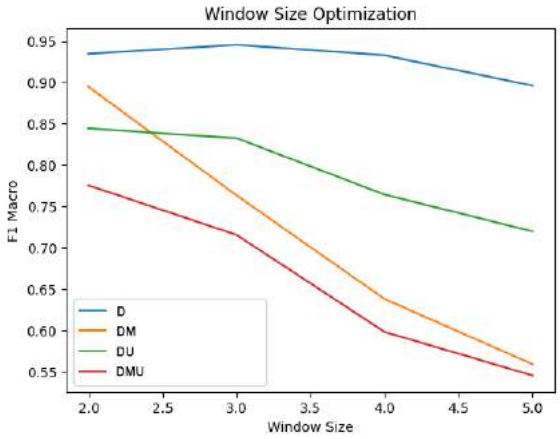
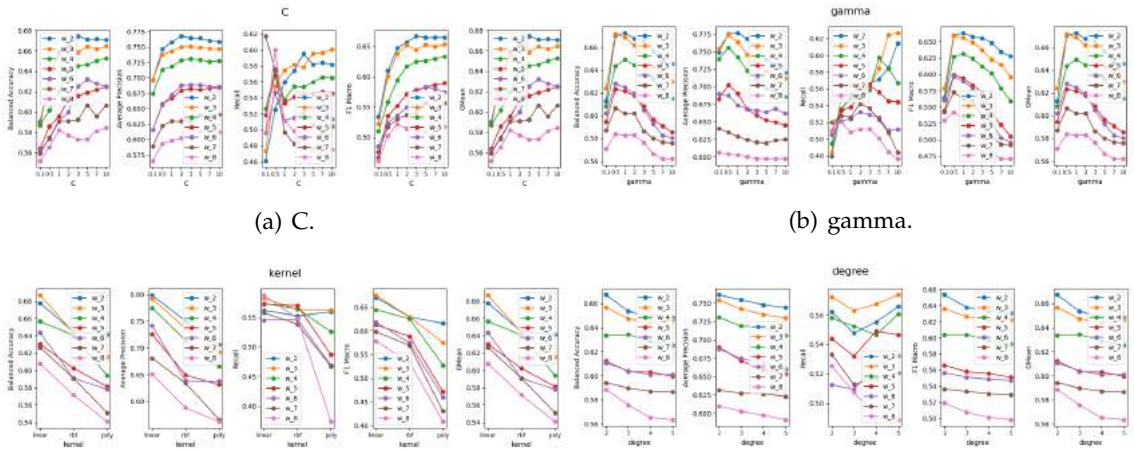


Figure B.38: L7c2 Porto de Mós Window Optimization for RF Upwelling Classification models.

B.1.2 Support Vector Machine

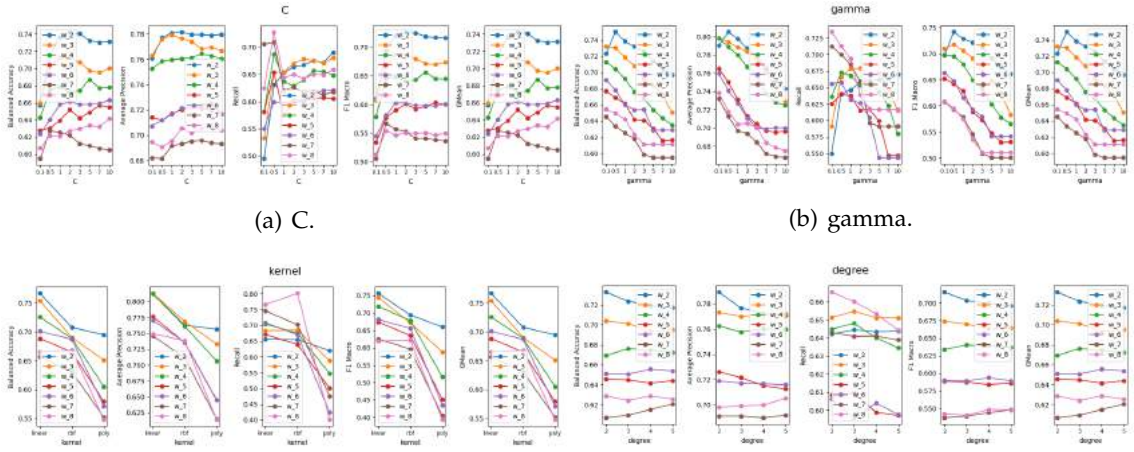
B.1.2.1 L1 Carreço SVM Classification



(a) C. (b) gamma.

(c) kernel. (d) degree.

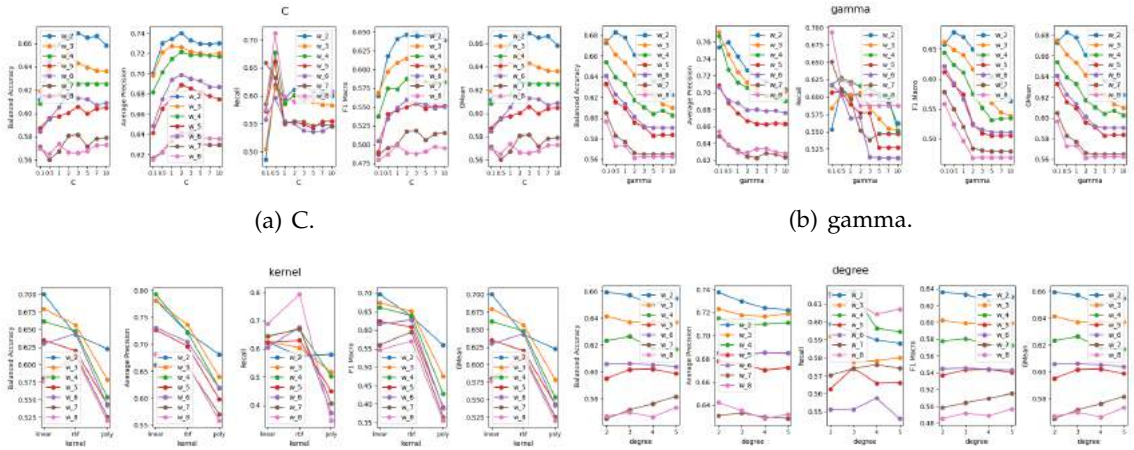
Figure B.39: SVM hyperparameters tuning for the L1-DM model.



(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.40: SVM hyperparameters tuning for the L1-DH model.

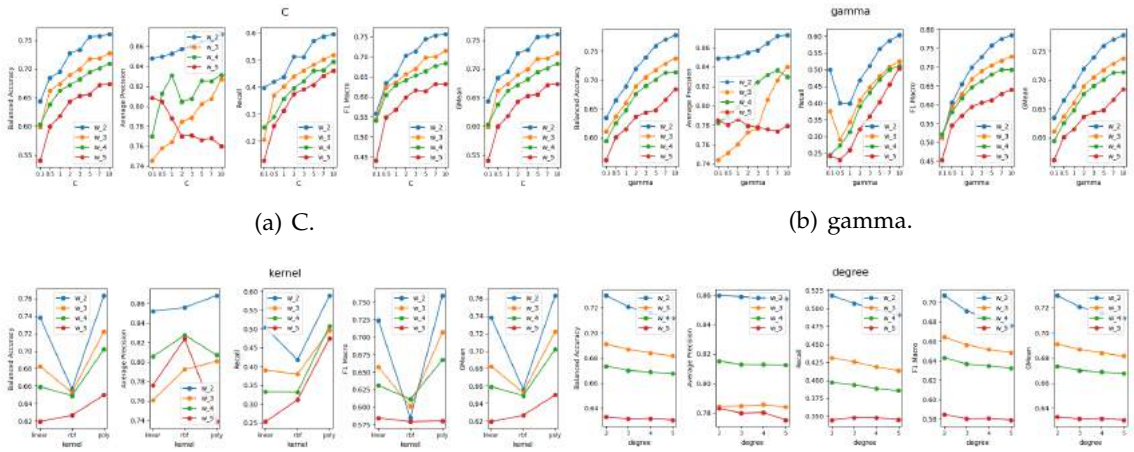


(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.41: SVM hyperparameters tuning for the L1-DMH model.

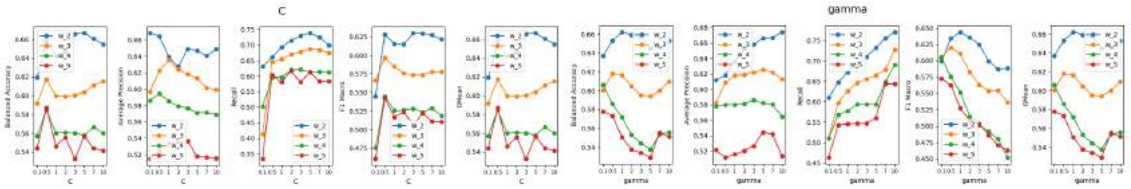
B.1.2.2 L1 Carreço SVM Upwelling Classification



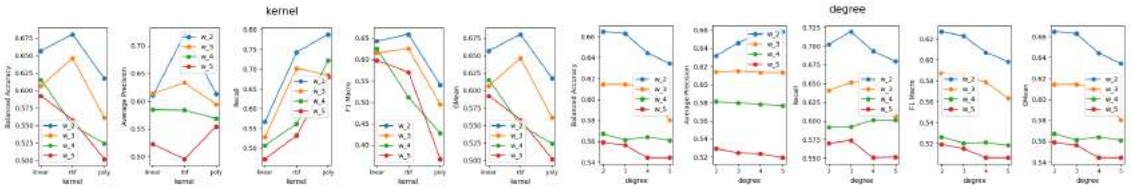
(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.42: SVM hyperparameters tuning for the L1-UP-D model.

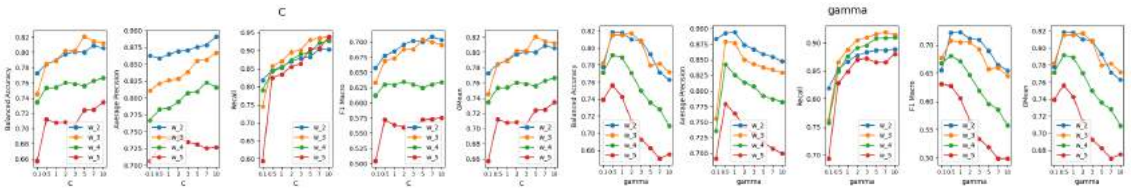


(a) C. (b) gamma.

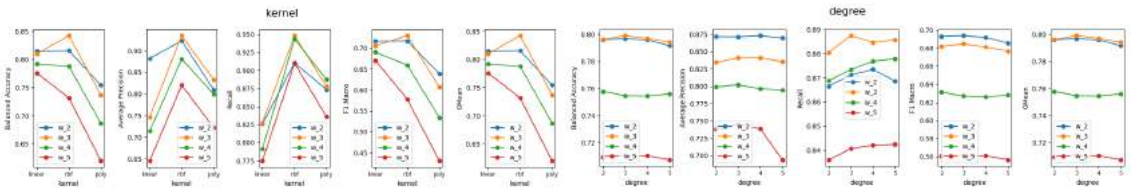


(c) kernel. (d) degree.

Figure B.43: SVM hyperparameters tuning for the L1-UP-DU model.

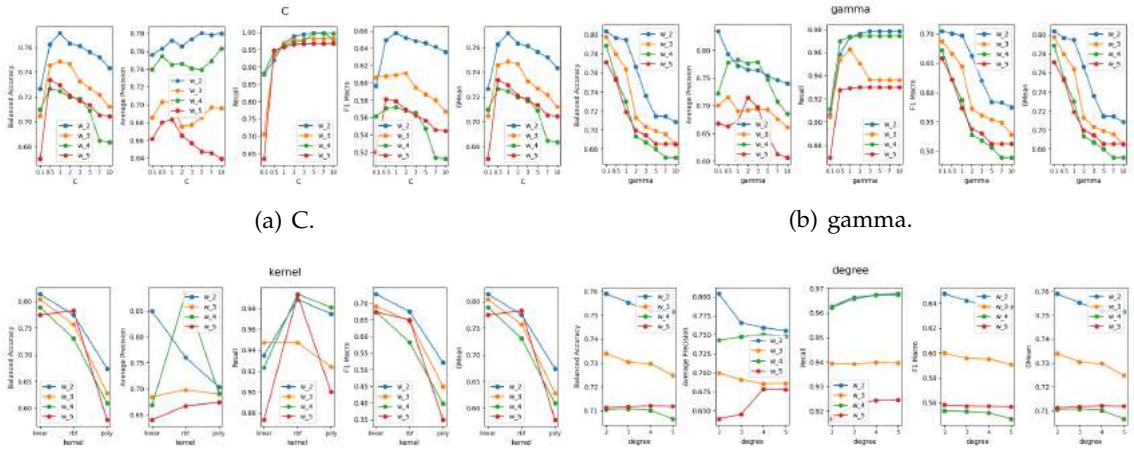


(a) C. (b) gamma.



(c) kernel. (d) degree.

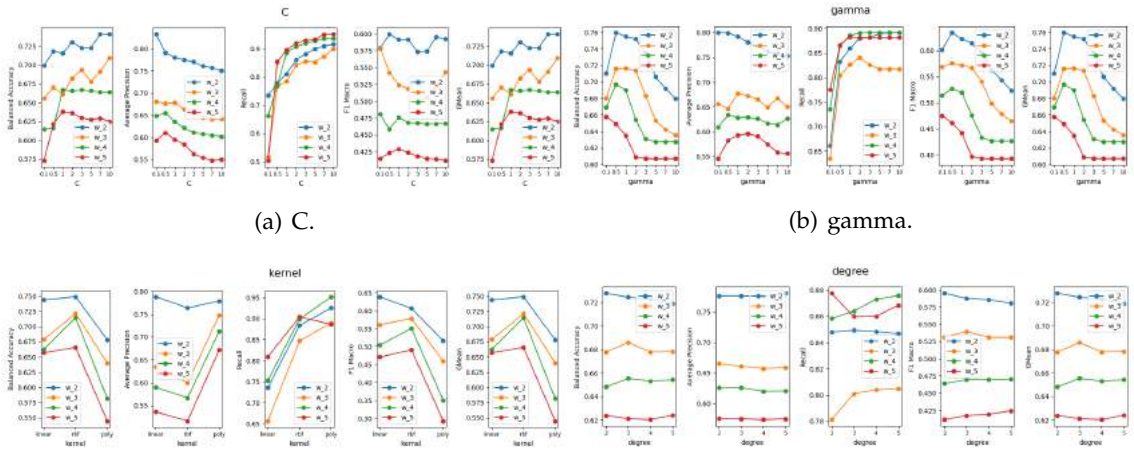
Figure B.44: SVM hyperparameters tuning for the L1-UP-DM model.



(a) C. (b) gamma.

(c) kernel. (d) degree.

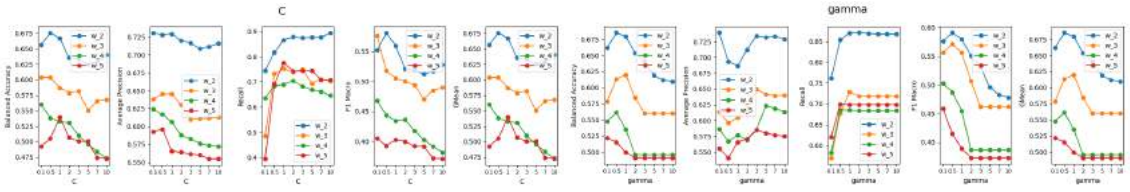
Figure B.45: SVM hyperparameters tuning for the L1-UP-DMU model.



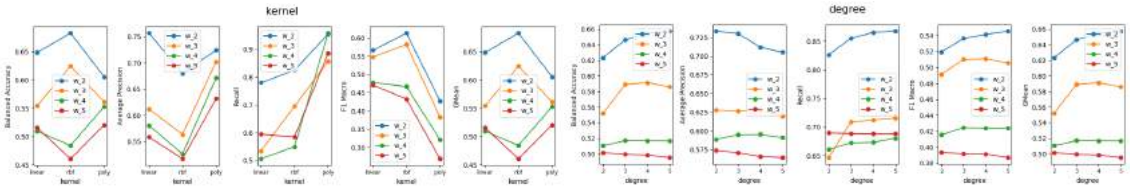
(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.46: SVM hyperparameters tuning for the L1-UP-DH model.

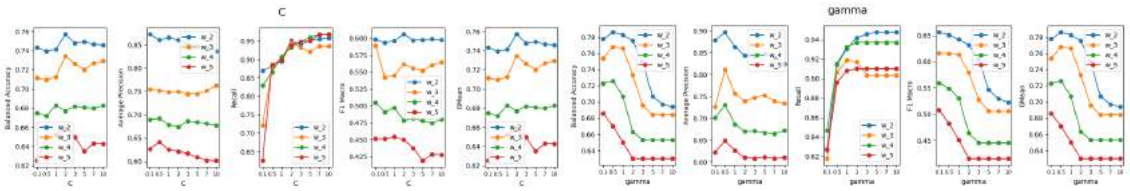


(a) C. (b) gamma.

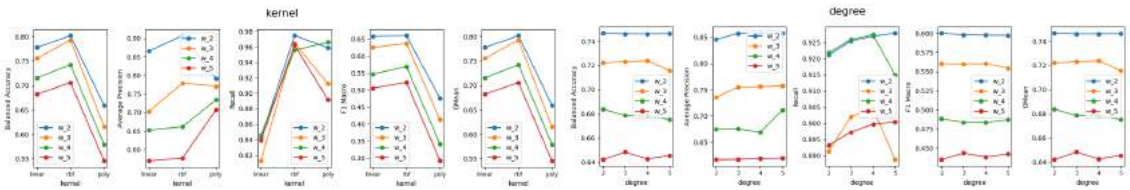


(c) kernel. (d) degree.

Figure B.47: SVM hyperparameters tuning for the L1-UP-DHU model.

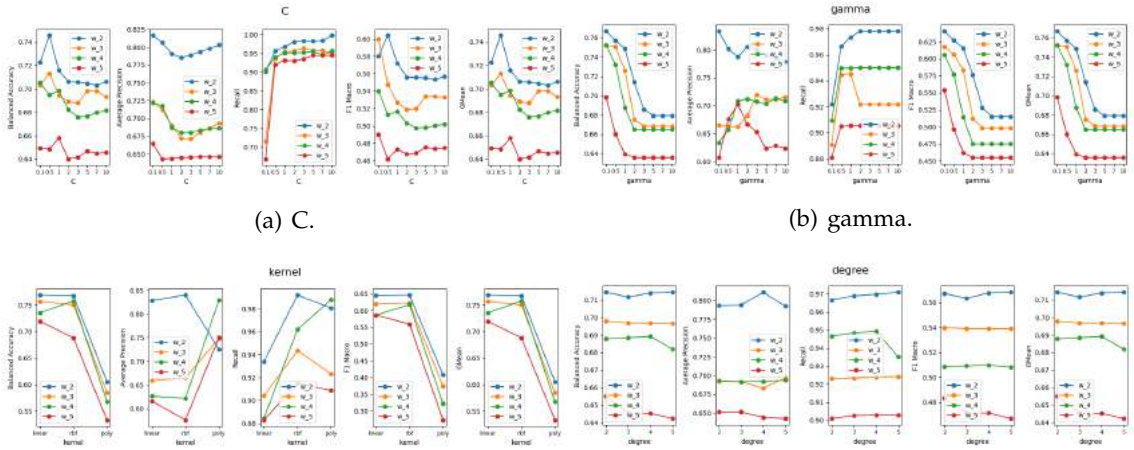


(a) C. (b) gamma.



(c) kernel. (d) degree.

Figure B.48: SVM hyperparameters tuning for the L1-UP-DMH model.



(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.49: SVM hyperparameters tuning for the L1-UP-DMHU model.

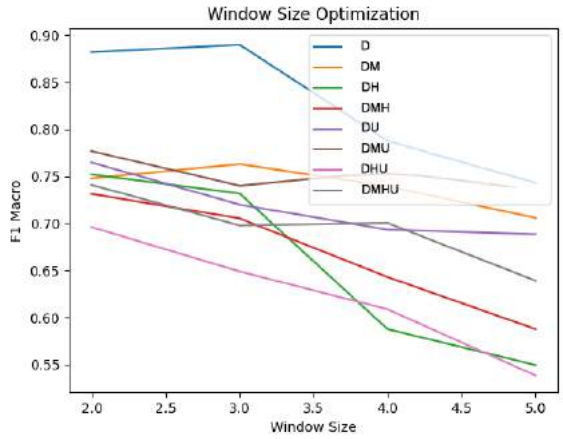
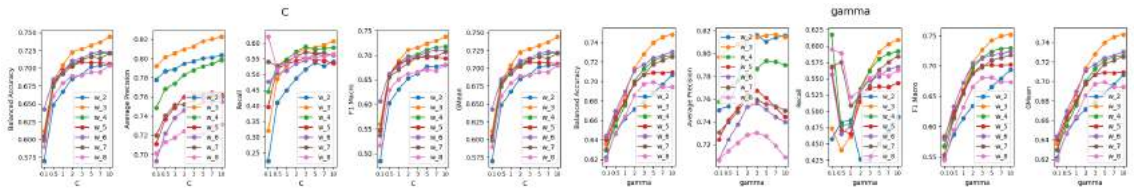
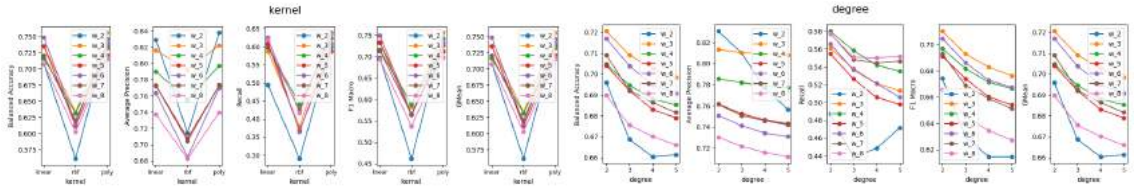


Figure B.50: L1 Carreço Window Optimization for SVM Upwelling Classification models.

B.1.2.3 L2 Leça da Palmeira SVM Classification

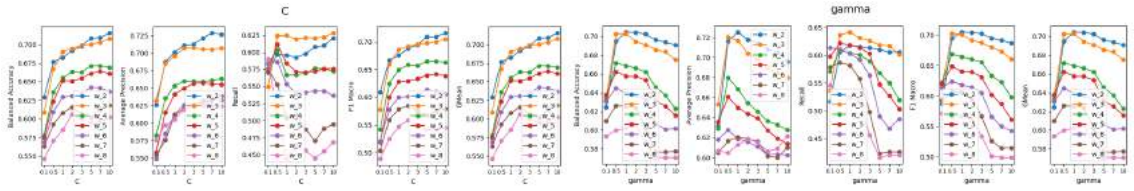


(a) C. (b) gamma.

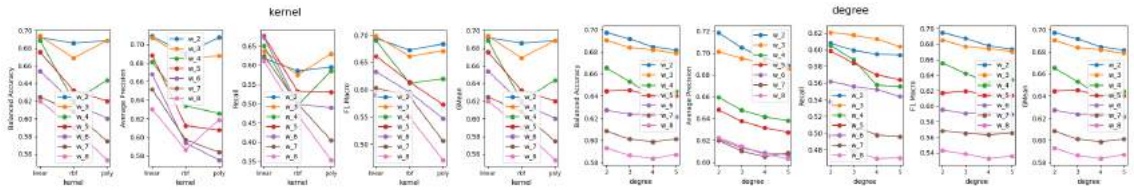


(c) kernel. (d) degree.

Figure B.51: SVM hyperparameters tuning for the L2-D model.



(a) C. (b) gamma.



(c) kernel. (d) degree.

Figure B.52: SVM hyperparameters tuning for the L2-DM model.

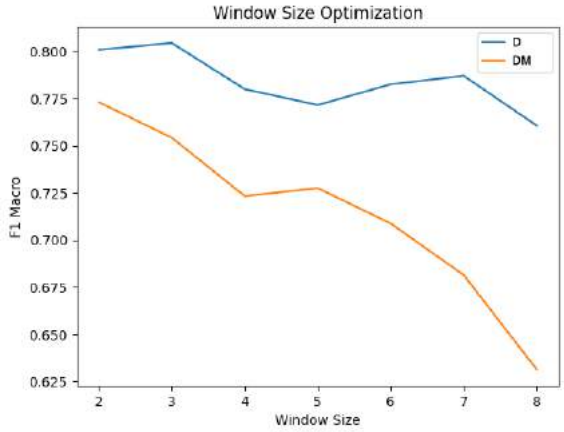
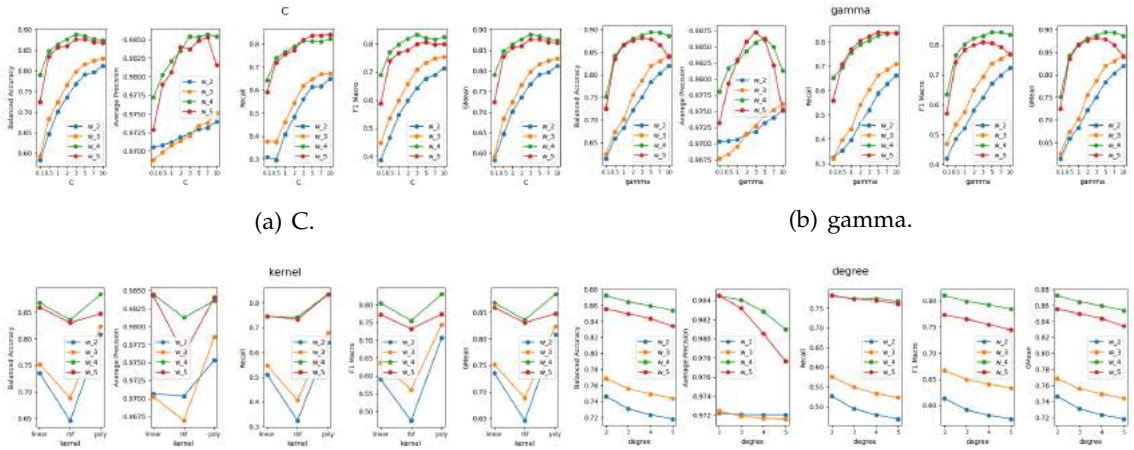


Figure B.53: L2 Leça da Palmeira Window Optimization for SVM Classification models.

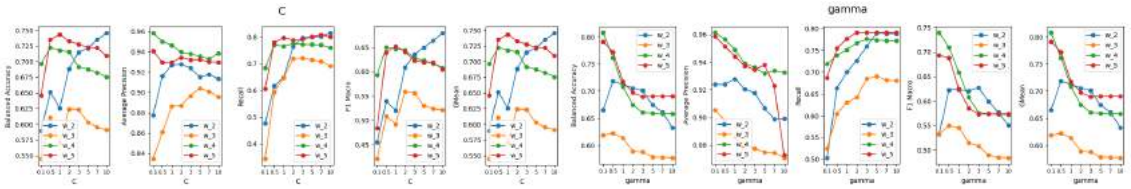
B.1.2.4 L2 Leça da Palmeira SVM Upwelling Classification



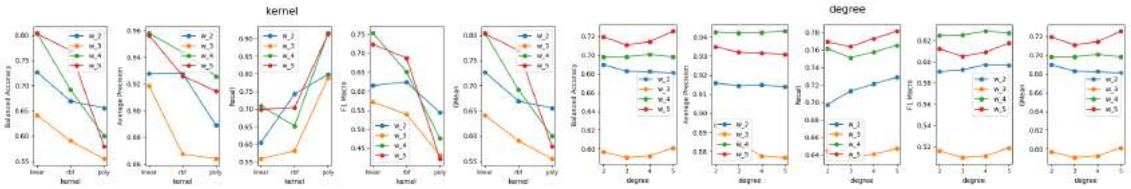
(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.54: SVM hyperparameters tuning for the L2-UP-D model.

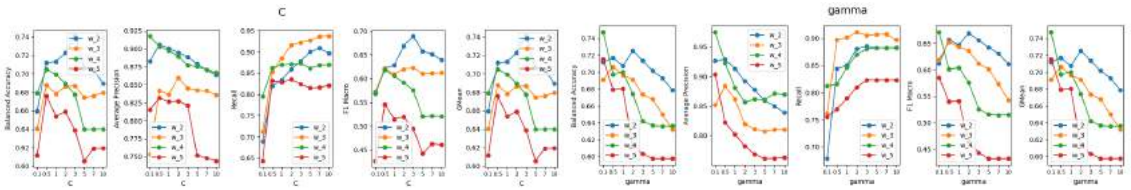


(a) C. (b) gamma.

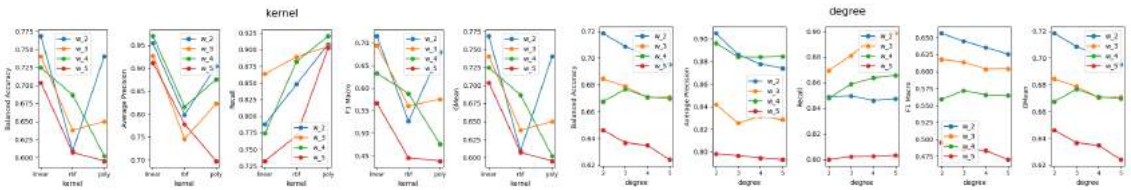


(c) kernel. (d) degree.

Figure B.55: SVM hyperparameters tuning for the L2-UP-DU model.

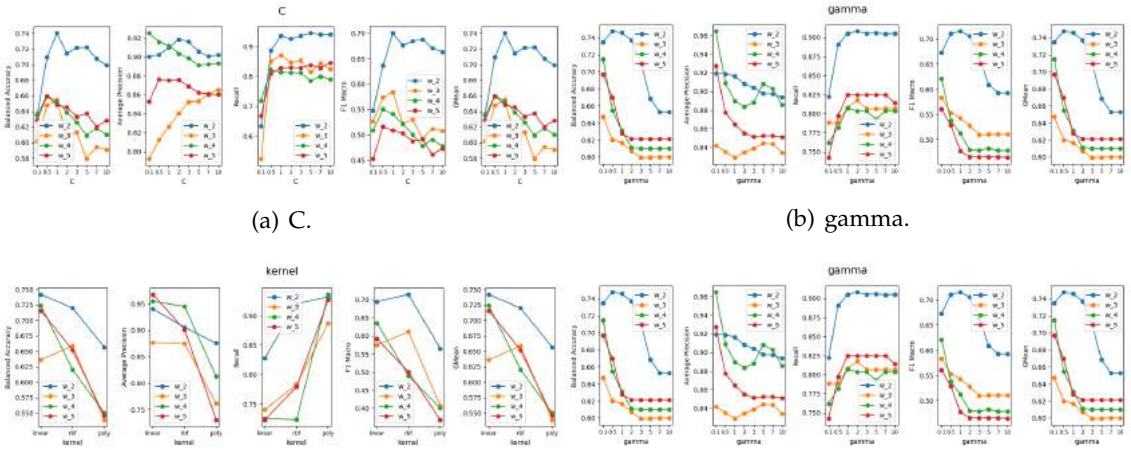


(a) C. (b) gamma.



(c) kernel. (d) degree.

Figure B.56: SVM hyperparameters tuning for the L2-UP-DM model.



(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.57: SVM hyperparameters tuning for the L2-UP-DMU model.

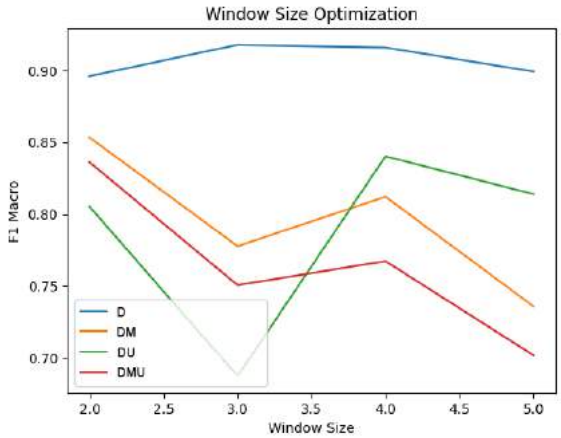
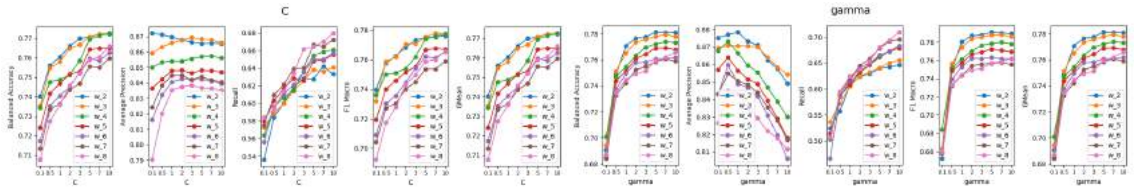
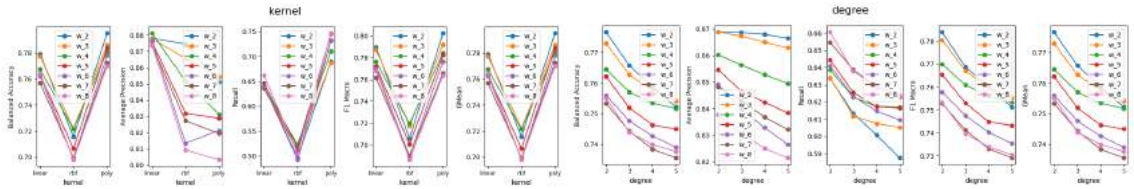


Figure B.58: L2 Leça da Palmeira Window Optimization for SVM Upwelling Classification models.

B.1.2.5 L5b Caparica SVM Classification

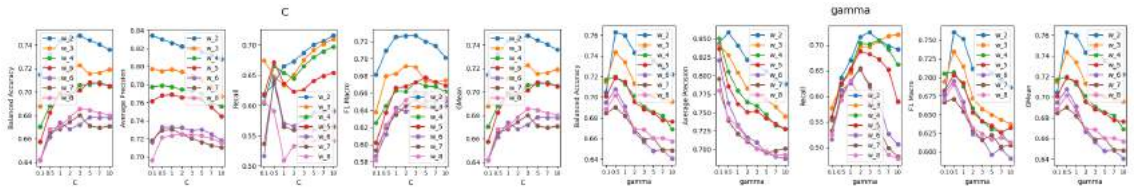


(a) C. (b) gamma.

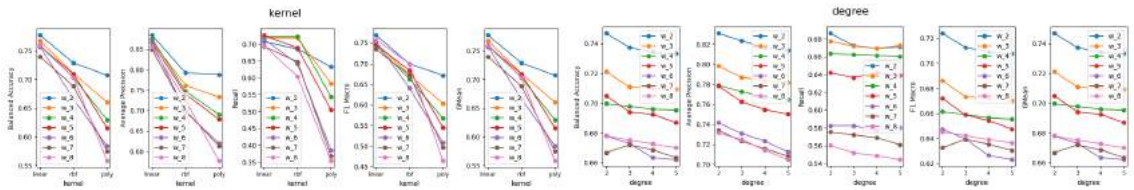


(c) kernel. (d) degree.

Figure B.59: SVM hyperparameters tuning for the L5b-D model.



(a) C. (b) gamma.



(c) kernel. (d) degree.

Figure B.60: SVR hyperparameters tuning for the L5b-DM model.

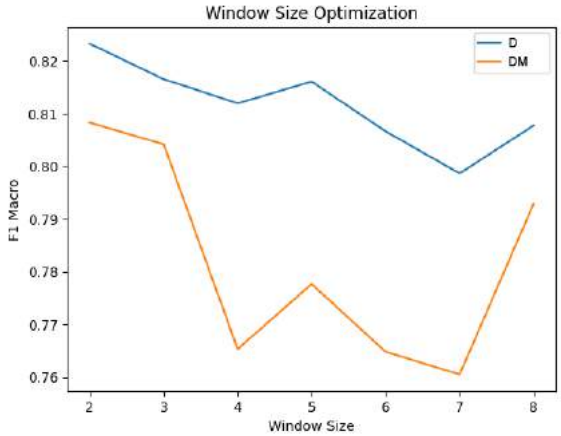
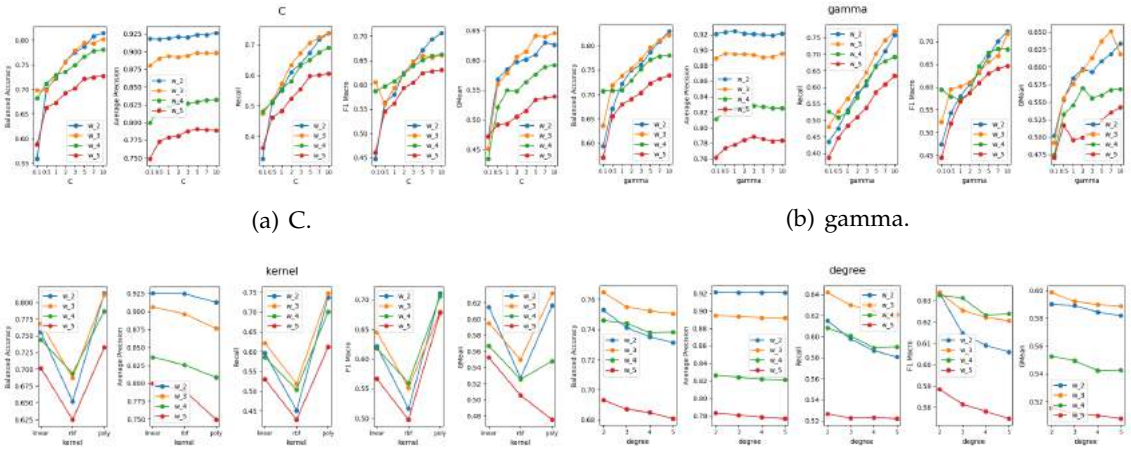


Figure B.61: L5b Caparica Window Optimization for SVM Classification models.

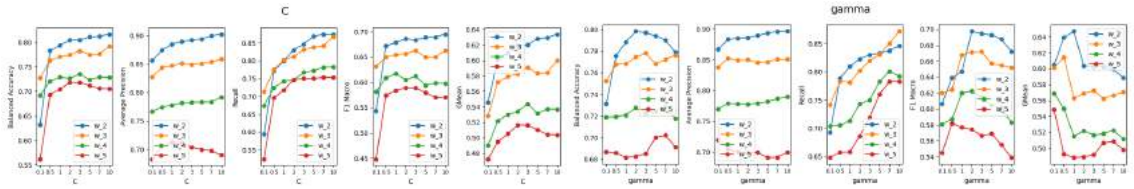
B.1.2.6 L5b Caparica SVM Upwelling Classification



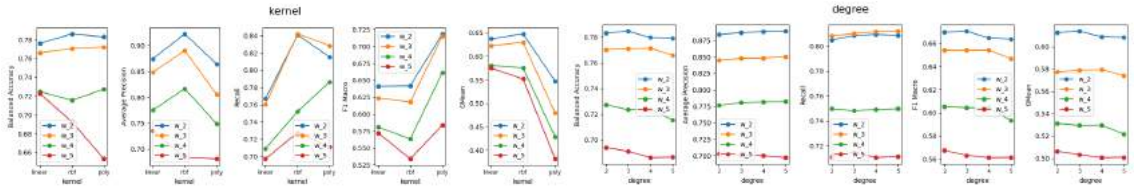
(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.62: SVM hyperparameters tuning for the L5b-UP-D model.

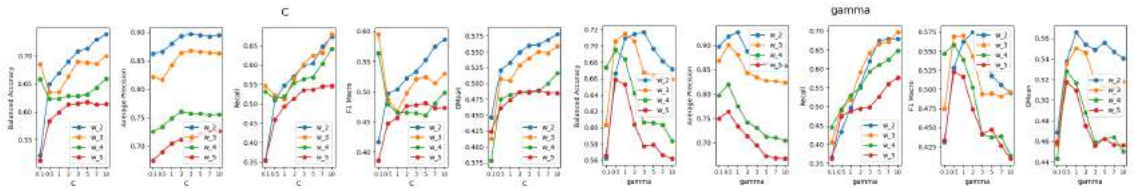


(a) C. (b) gamma.

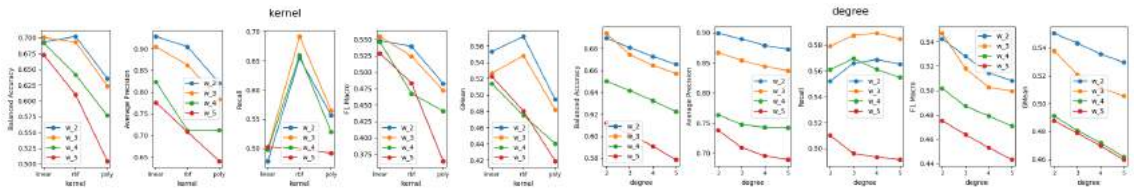


(c) kernel. (d) degree.

Figure B.63: SVM hyperparameters tuning for the L5b-UP-DU model.

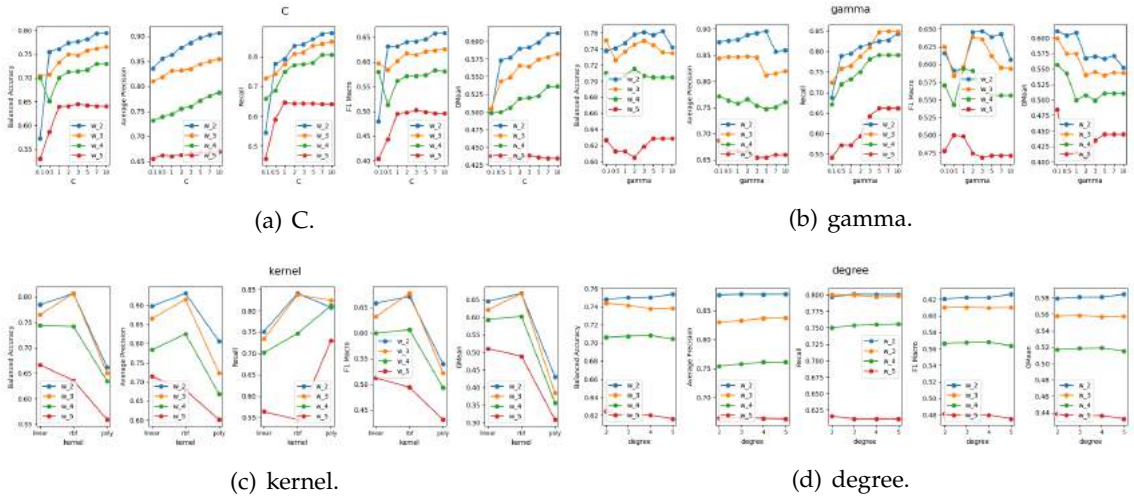


(a) C. (b) gamma.



(c) kernel. (d) degree.

Figure B.64: SVM hyperparameters tuning for the L5b-UP-DM model.



(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.65: SVM hyperparameters tuning for the L5b-UP-DMU model.

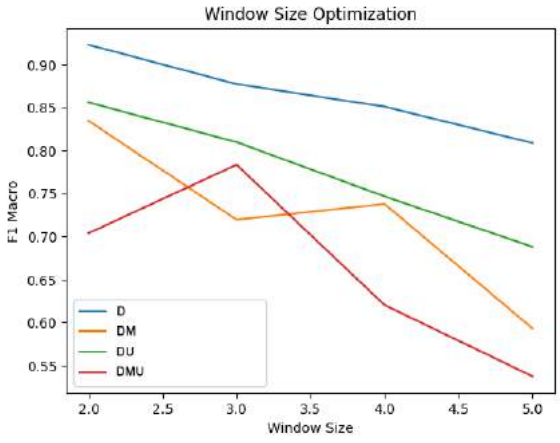
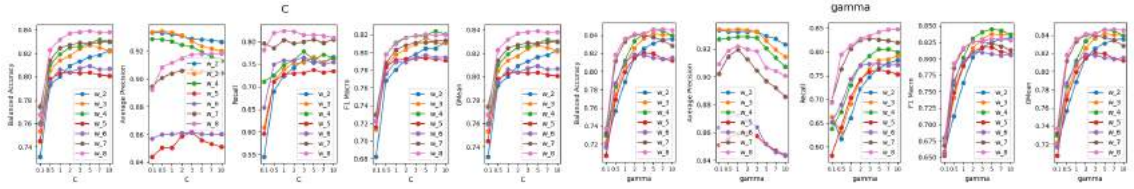
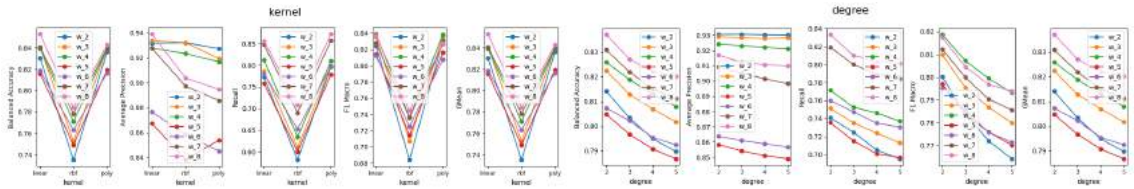


Figure B.66: L5b Caparica Window Optimization for SVM Classification Regression models.

B.1.2.7 RIAV1 Triângulo SVM Classification

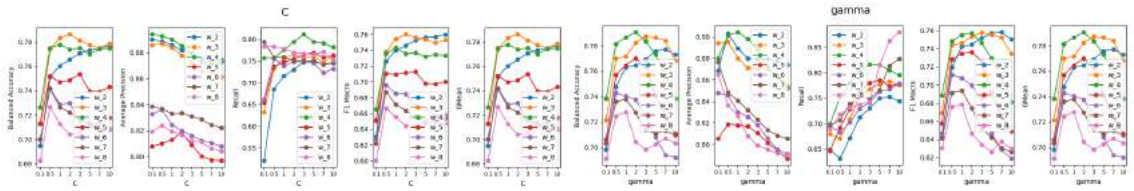


(a) C. (b) gamma.

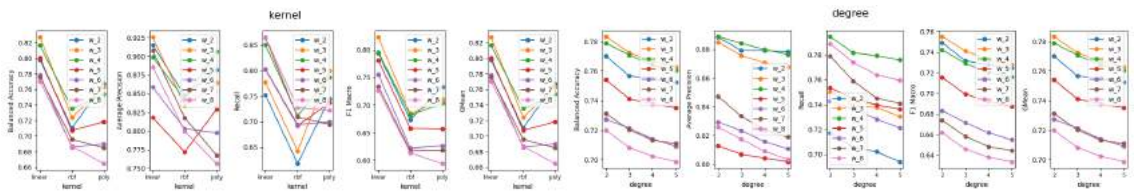


(c) kernel. (d) degree.

Figure B.67: SVM hyperparameters tuning for the RIAV1-D model.

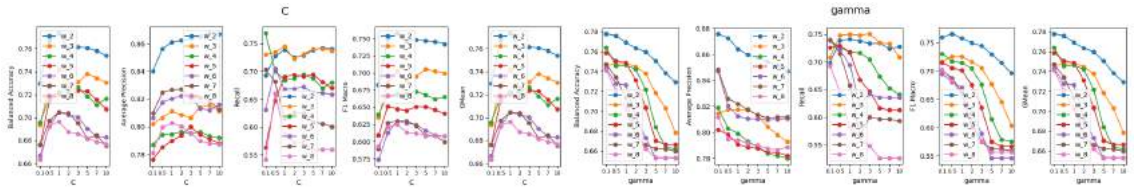


(a) C. (b) gamma.

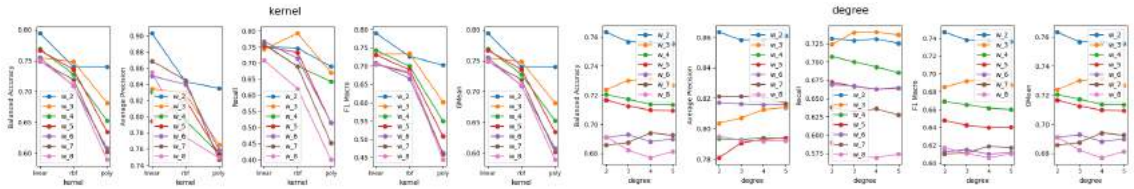


(c) kernel. (d) degree.

Figure B.68: SVM hyperparameters tuning for the RIAV1-DM model.

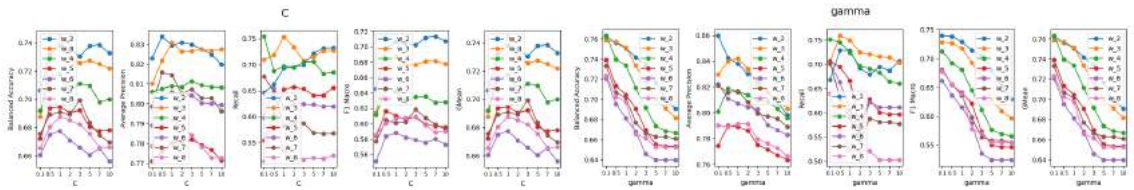


(a) C. (b) gamma.

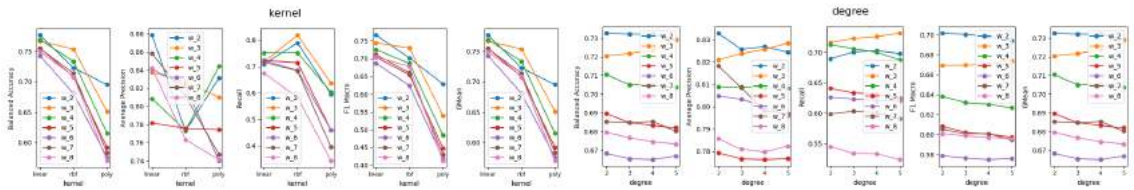


(c) kernel. (d) degree.

Figure B.69: SVM hyperparameters tuning for the RIAV1-DH model.



(a) C. (b) gamma.



(c) kernel. (d) degree.

Figure B.70: SVM hyperparameters tuning for the RIAV1-DMH model.

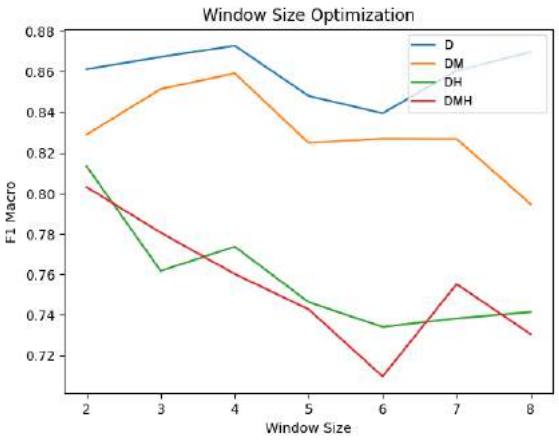
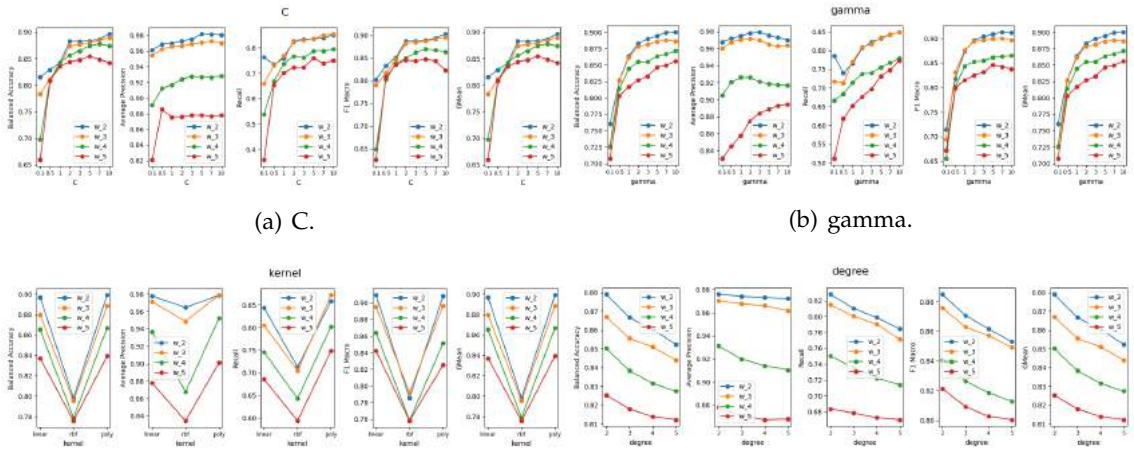


Figure B.71: RIAV1 Triângulo Window Optimization for SVM Classification models.

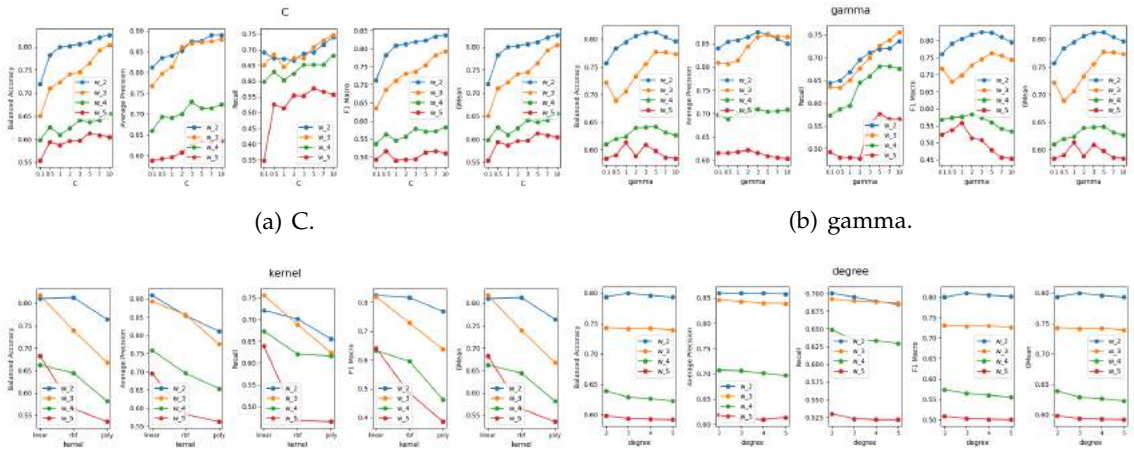
B.1.2.8 L7c2 Porto de Mós SVM Upwelling Classification



(a) C. (b) gamma.

(c) kernel. (d) degree.

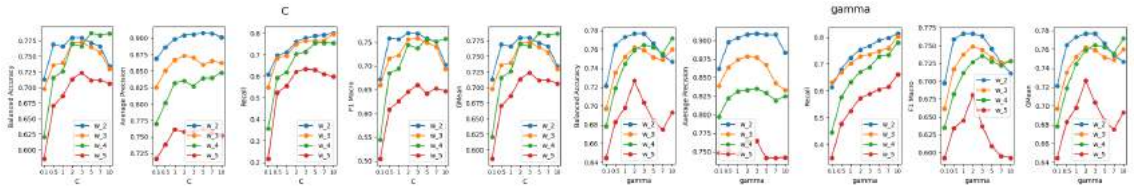
Figure B.72: SVM hyperparameters tuning for the L7c2-UP-D model.



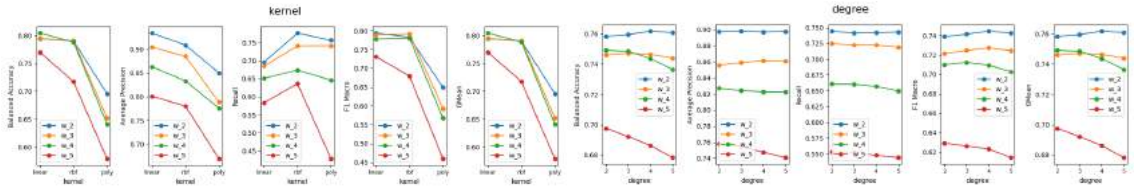
(a) C. (b) gamma.

(c) kernel. (d) degree.

Figure B.73: SVM hyperparameters tuning for the L7c2-UP-DU model.

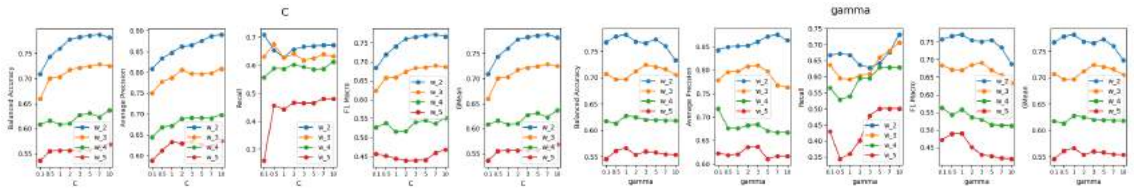


(a) C. (b) gamma.

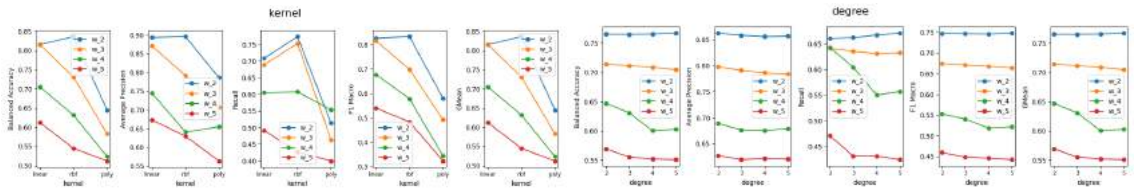


(c) kernel. (d) degree.

Figure B.74: SVM hyperparameters tuning for the L7c2-UP-DM model.



(a) C. (b) gamma.



(c) kernel. (d) degree.

Figure B.75: SVM hyperparameters tuning for the L7c2-UP-DMU model.

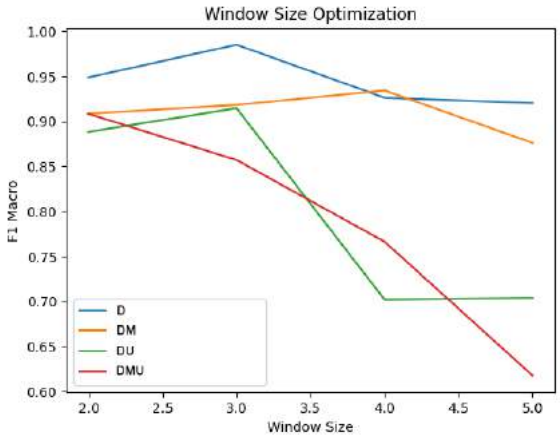
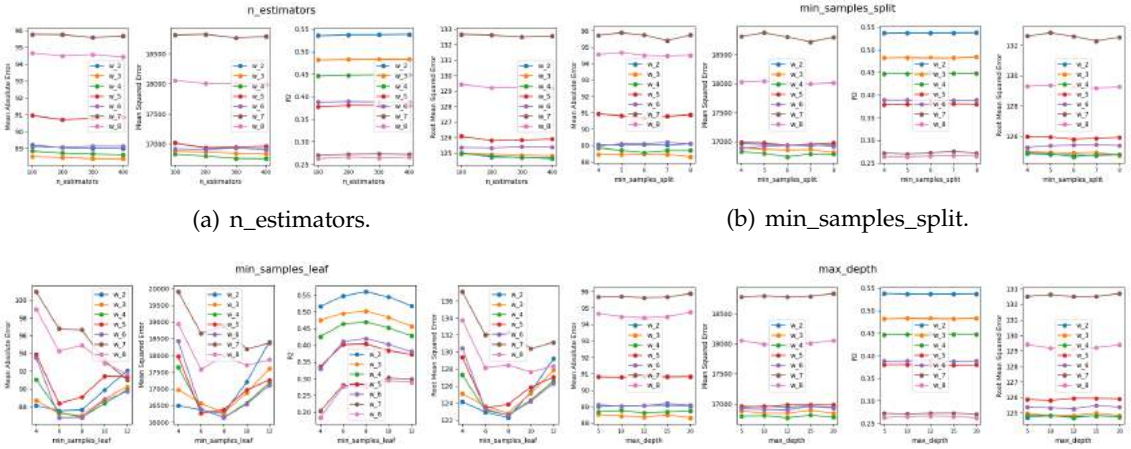


Figure B.76: L7c2 Porto de Mós Window Optimization for SVM Upwelling Classification models.

B.2 Regression

B.2.1 Random Forest

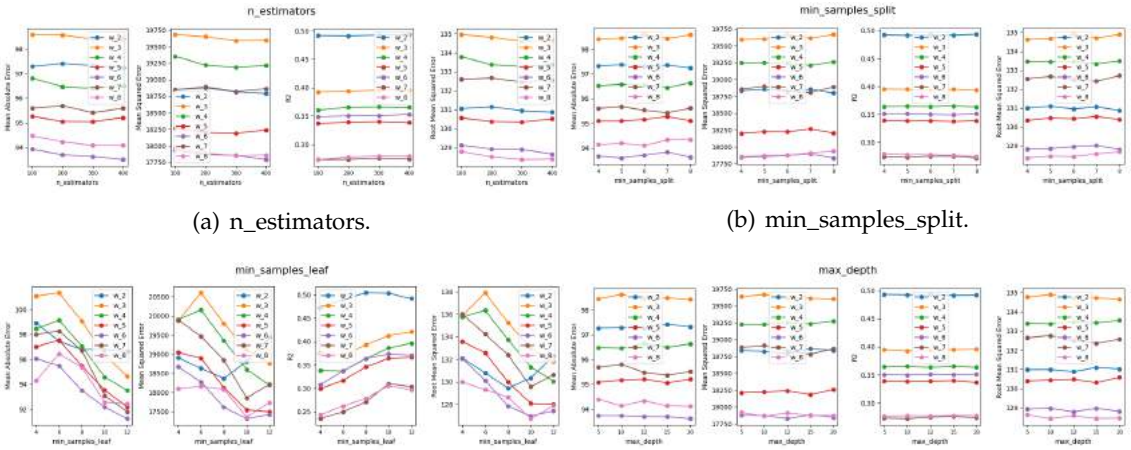
B.2.1.1 L1 Carreço RF Regression



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

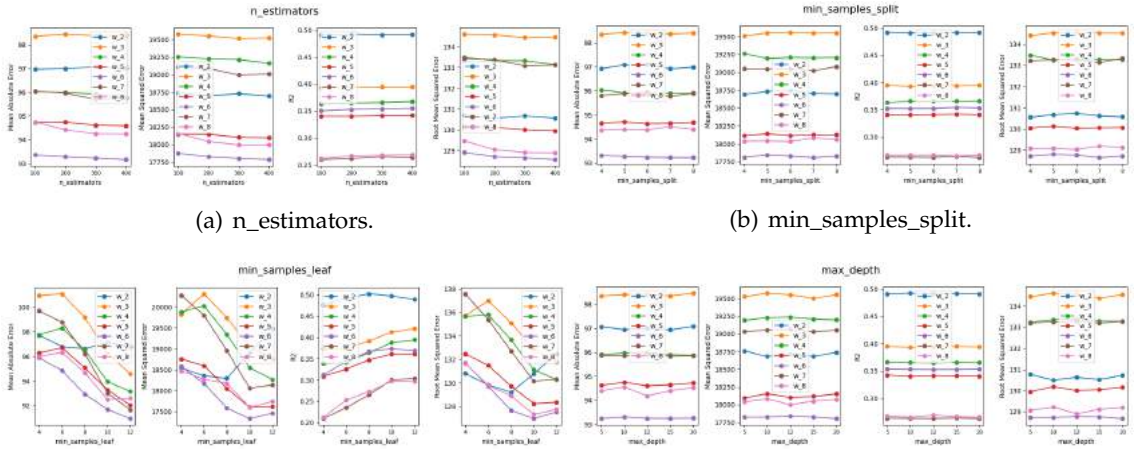
Figure B.77: RF hyperparameters tuning for the L1-DM model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.78: RF hyperparameters tuning for the L1-DH model.

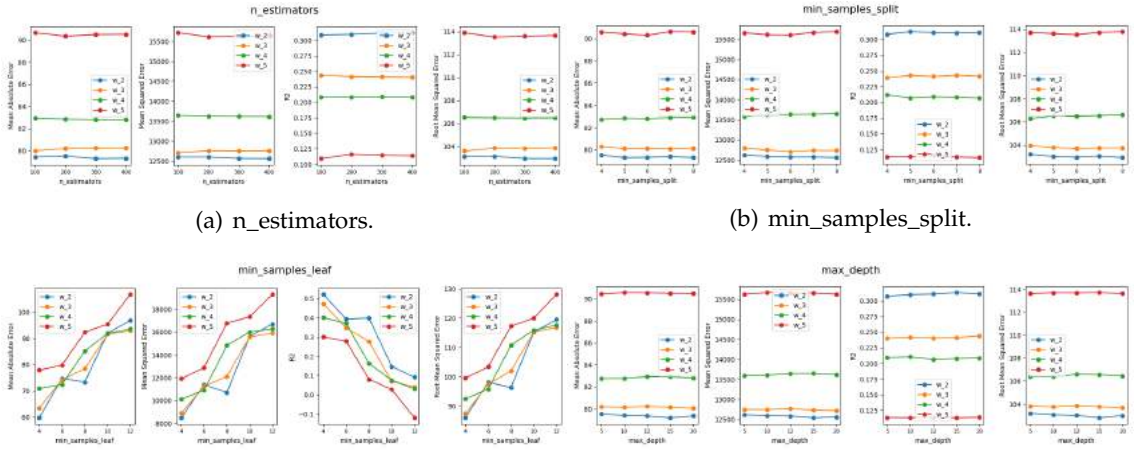


(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.79: RF hyperparameters tuning for the L1-DMH model.

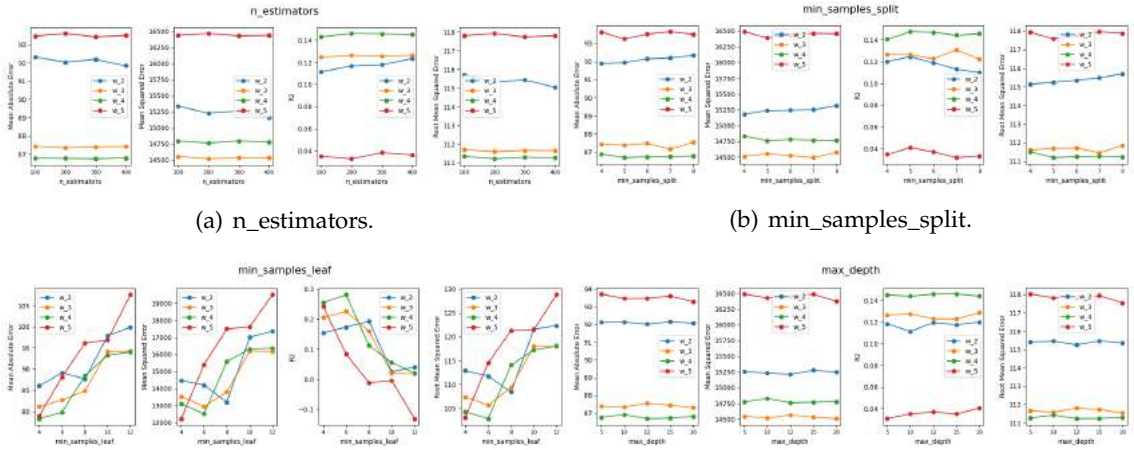
B.2.1.2 L1 Carreço RF Upwelling Regression



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

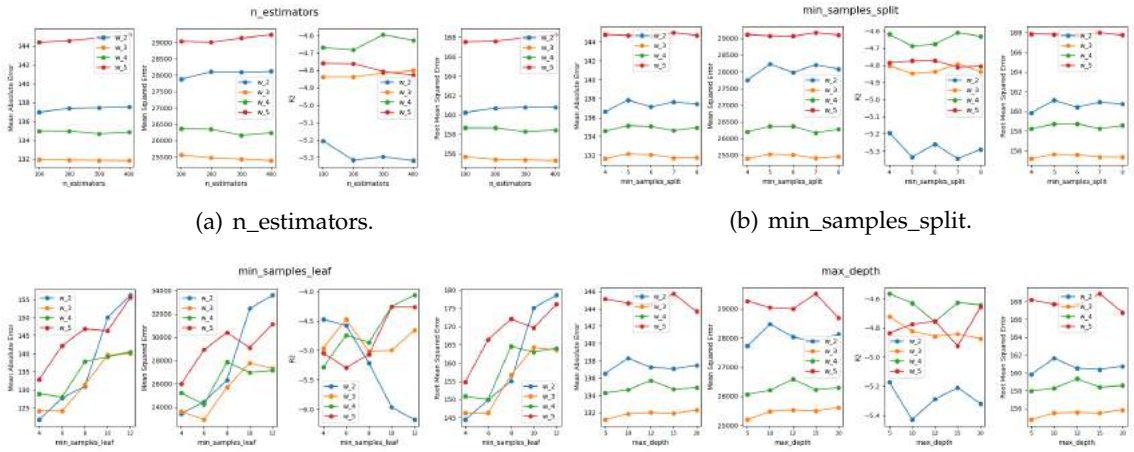
Figure B.80: RF hyperparameters tuning for the L1-UP-D model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

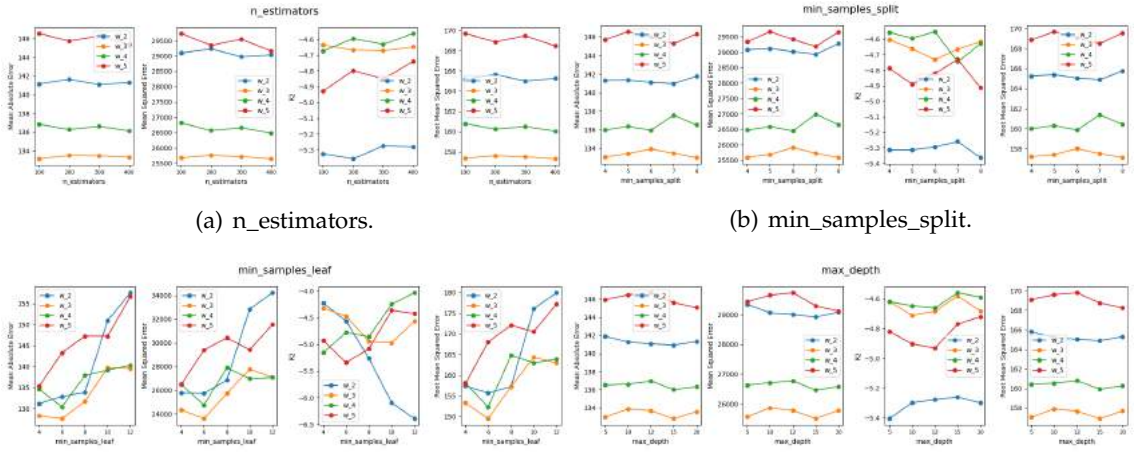
Figure B.81: RF hyperparameters tuning for the L1-UP-DU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

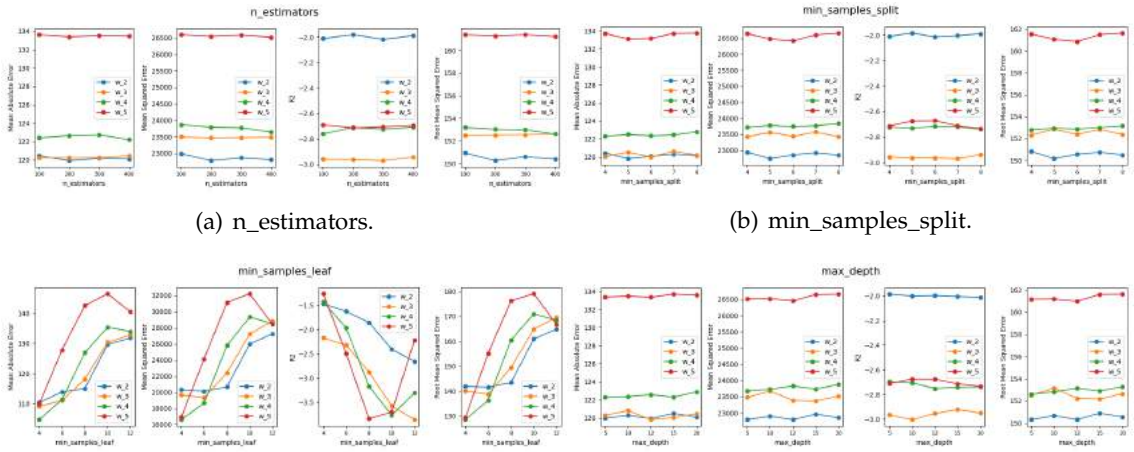
Figure B.82: RF hyperparameters tuning for the L1-UP-DM model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

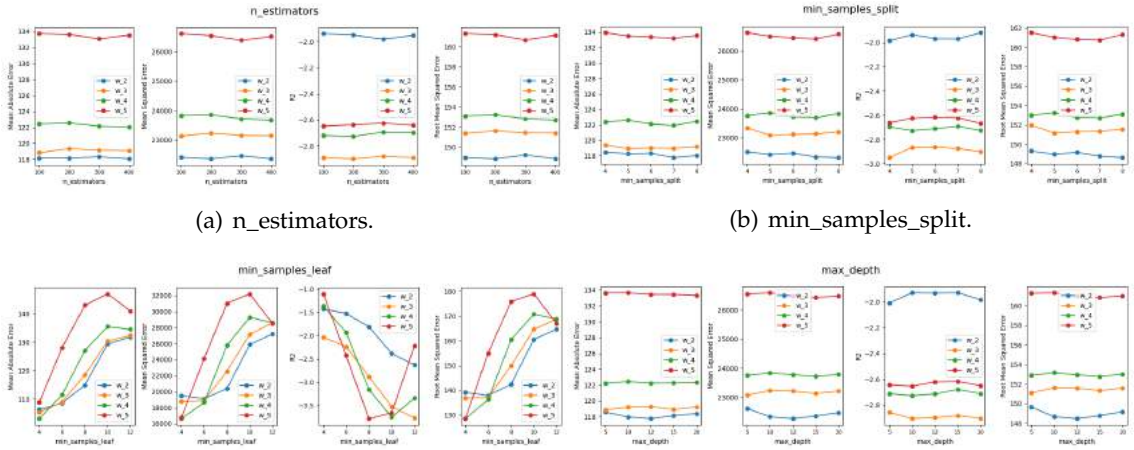
Figure B.83: RF hyperparameters tuning for the L1-UP-DMU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

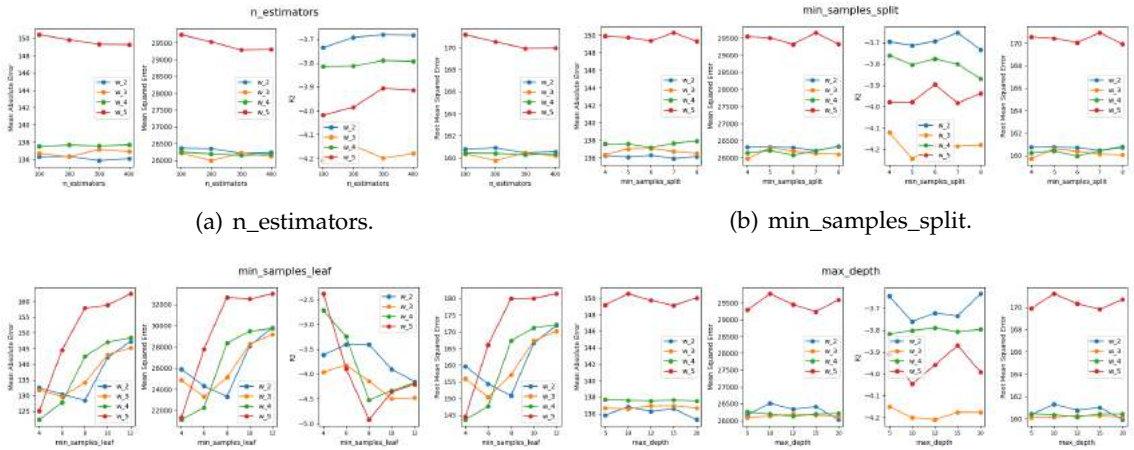
Figure B.84: RF hyperparameters tuning for the L1-UP-DH model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

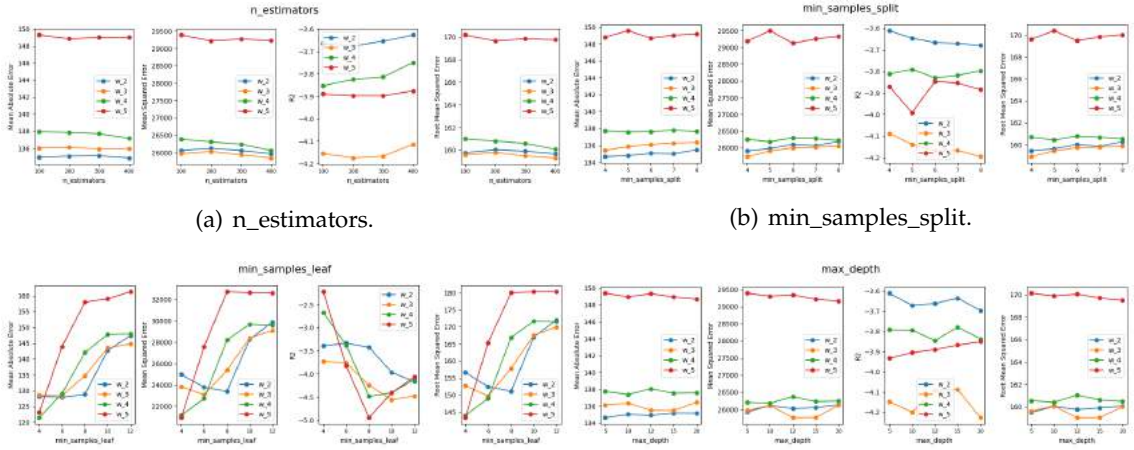
Figure B.85: RF hyperparameters tuning for the L1-UP-DHU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.86: RF hyperparameters tuning for the L1-UP-DMH model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.87: RF hyperparameters tuning for the L1-UP-DMHU model.

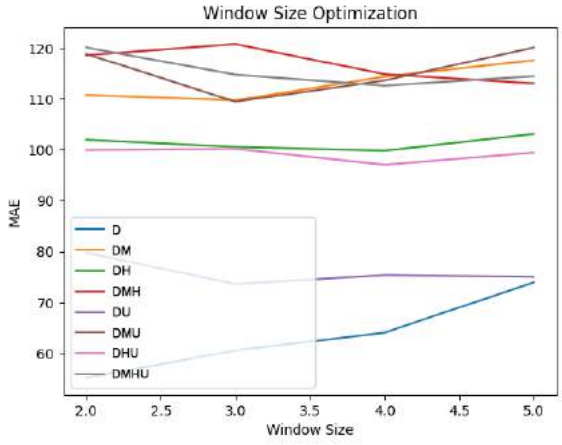
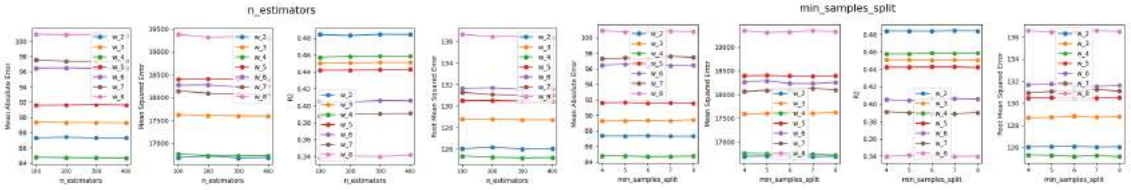


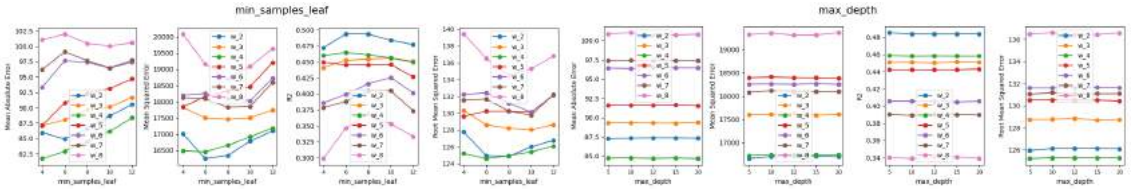
Figure B.88: L1 Carreço Window Optimization for RF Regression models.

B.2.1.3 L2 Leça da Palmeira RF Regression



(a) n_estimators.

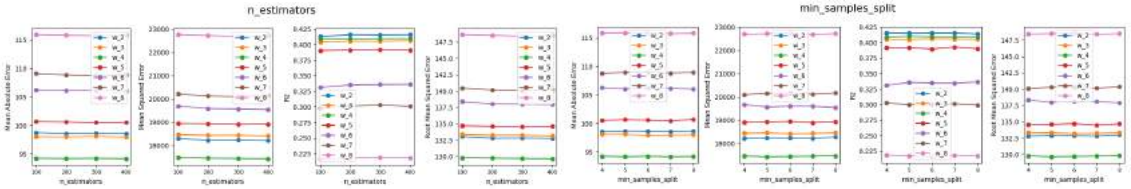
(b) min_samples_split.



(c) min_samples_leaf.

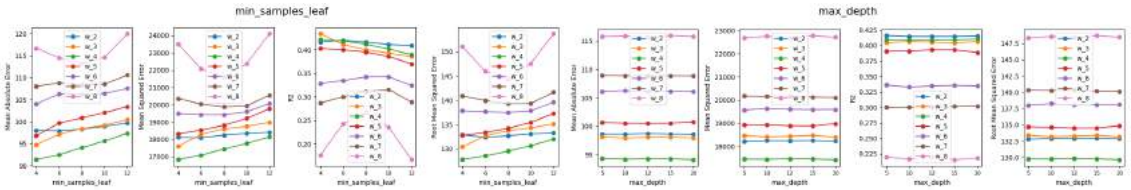
(d) max_depth.

Figure B.89: RF hyperparameters tuning for the L2-D model.



(a) n_estimators.

(b) min_samples_split.



(c) min_samples_leaf.

(d) max_depth.

Figure B.90: RF hyperparameters tuning for the L2-DM model.

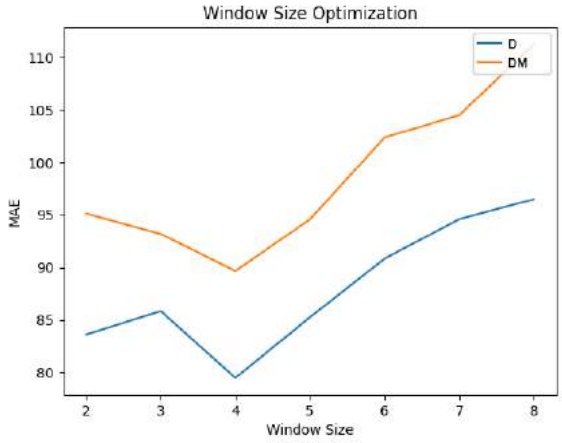
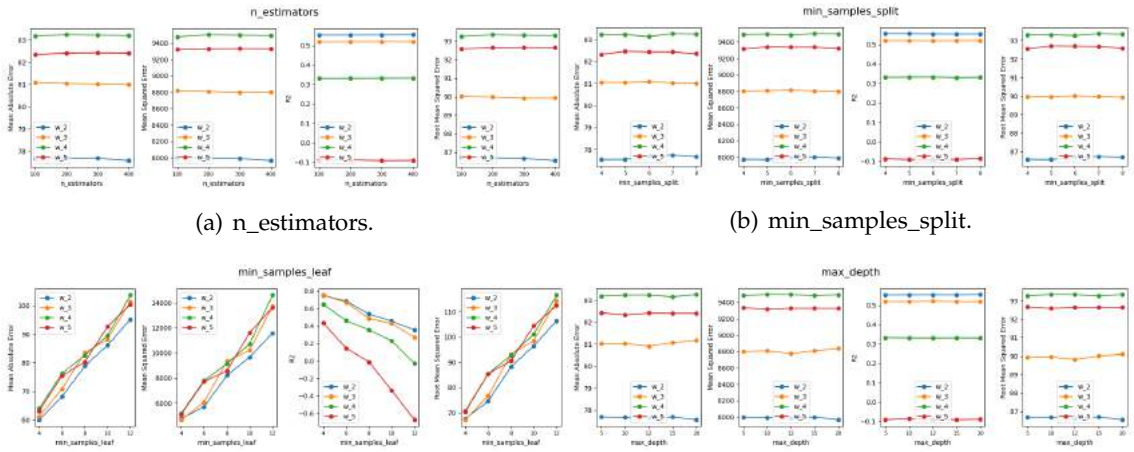


Figure B.91: L2 Leça da Palmeira Window Optimization for RF Regression models.

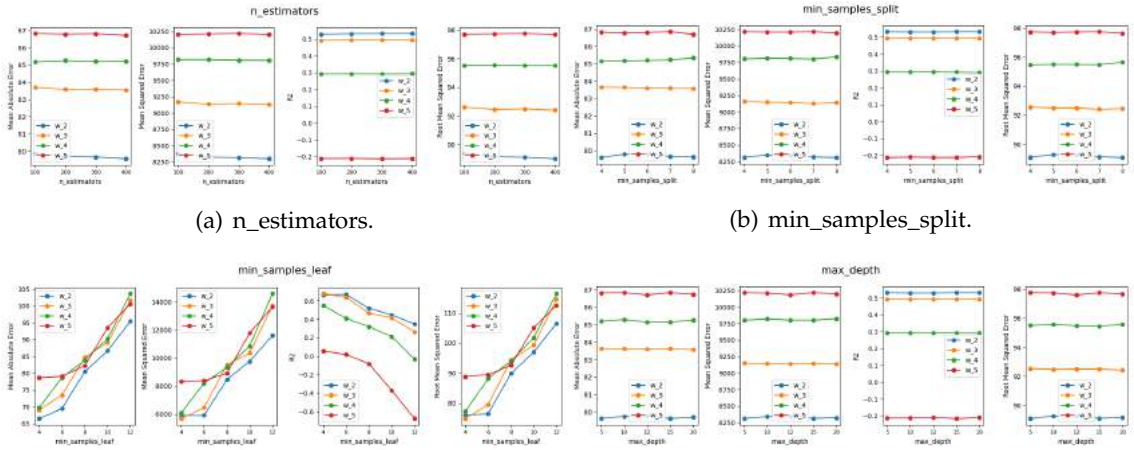
B.2.1.4 L2 Leça da Palmeira RF Upwelling Regression



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

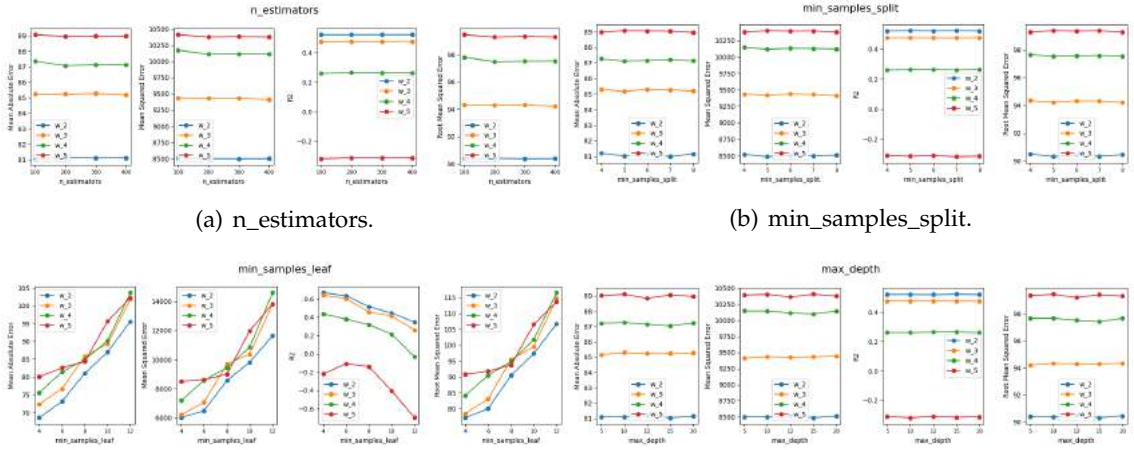
Figure B.92: RF hyperparameters tuning for the L2-UP-D model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

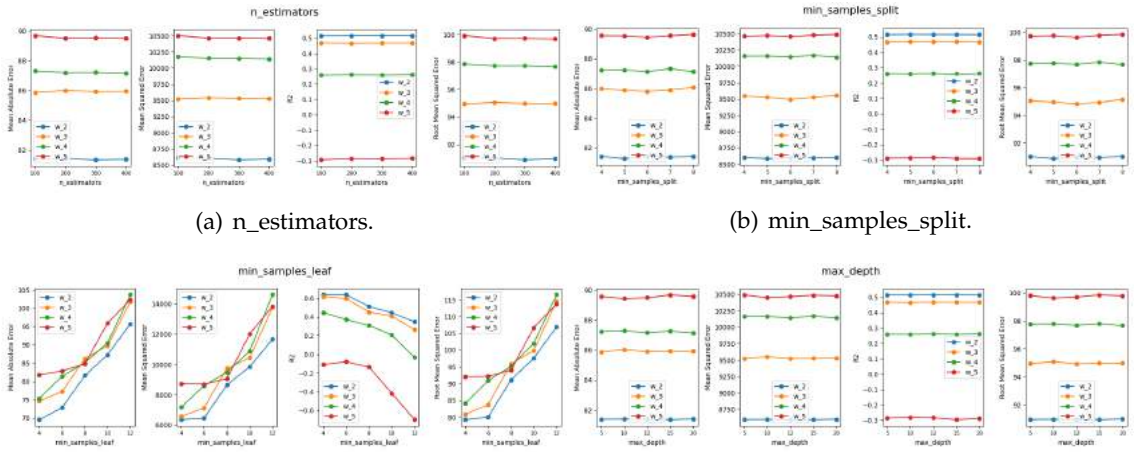
Figure B.93: RF hyperparameters tuning for the L2-UP-DU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.94: RF hyperparameters tuning for the L2-UP-DM model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.95: RF hyperparameters tuning for the L2-UP-DMU model.

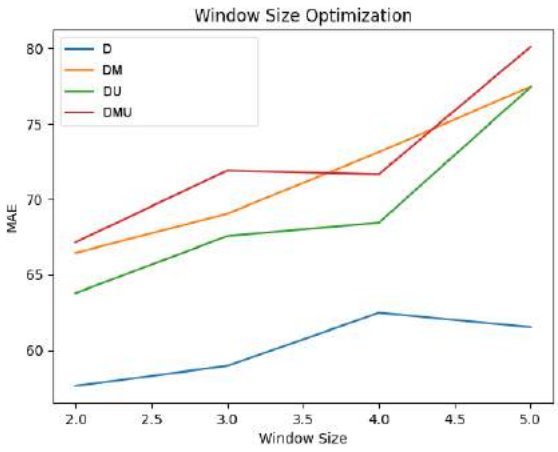
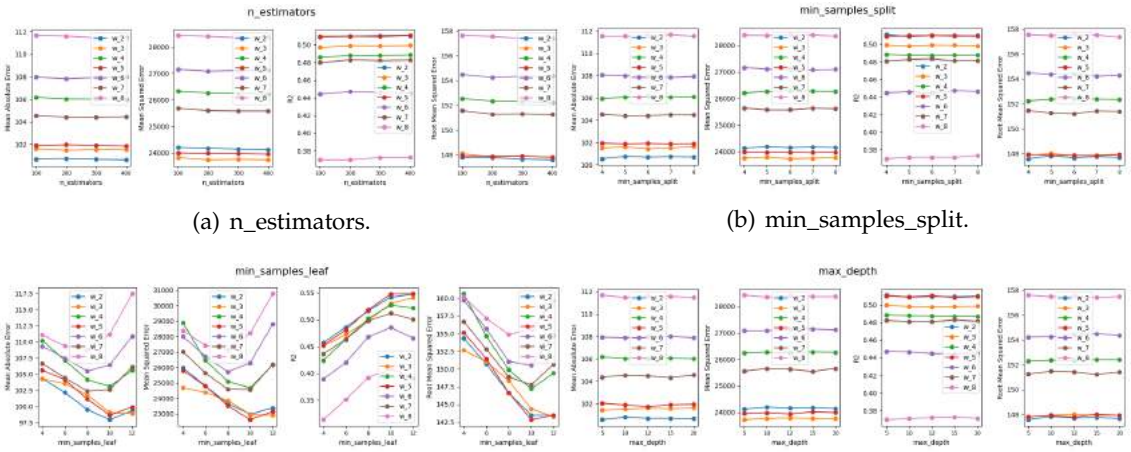


Figure B.96: L2 Leça da Palmeira Window Optimization for RF Upwelling Regression models.

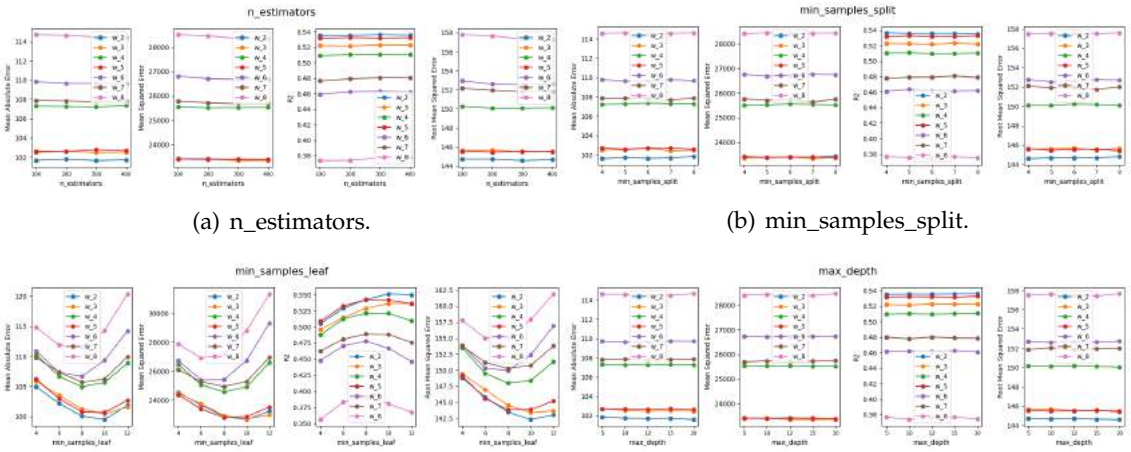
B.2.1.5 L5b Caparica RF Regression



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.97: RF hyperparameters tuning for the L5b-D model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.98: RF hyperparameters tuning for the L5b-DM model.

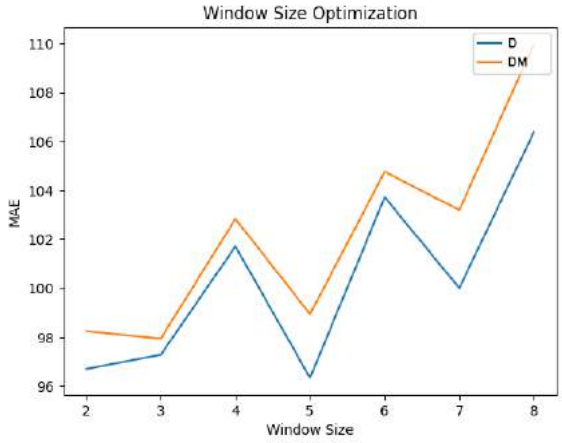
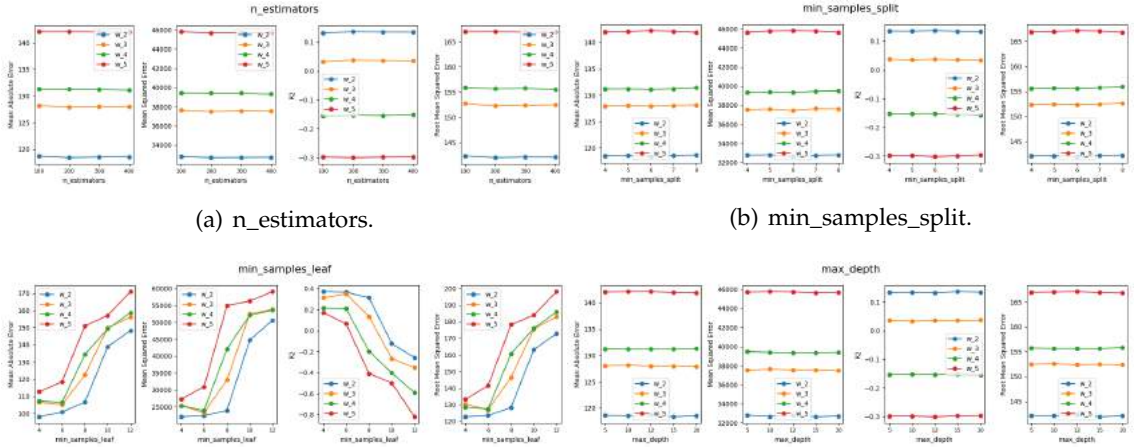


Figure B.99: L5b Caparica Window Optimization for RF Regression models.

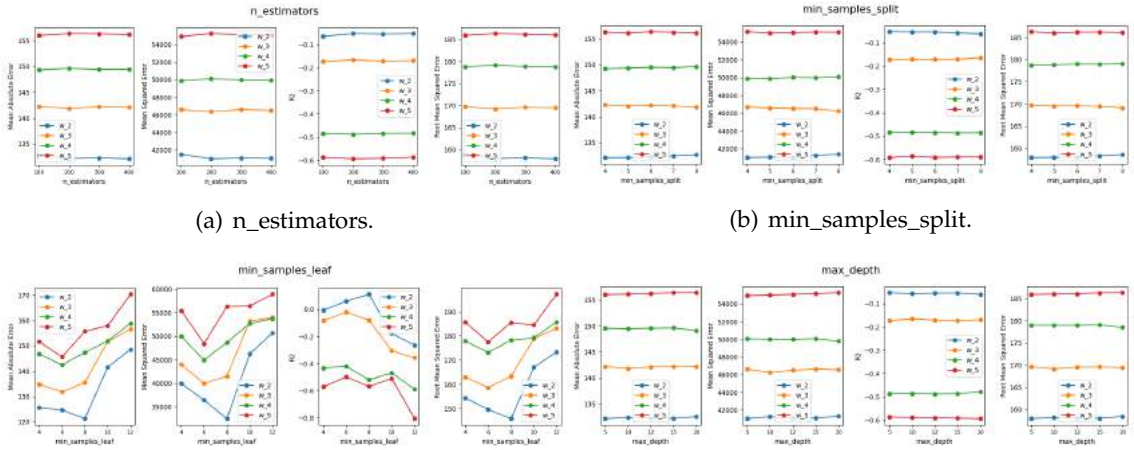
B.2.1.6 L5b Caparica RF Upwelling Regression



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

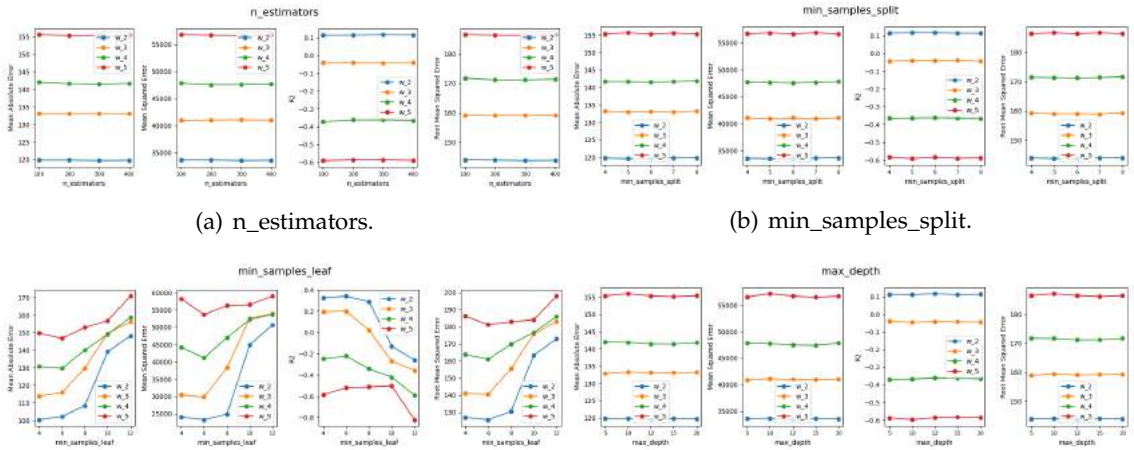
Figure B.100: RF hyperparameters tuning for the L5b-UP-D model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

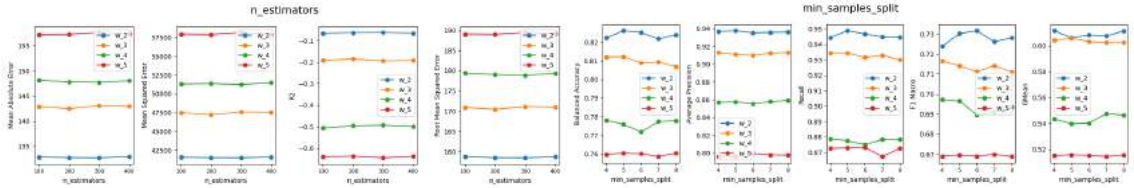
Figure B.101: RF hyperparameters tuning for the L5b-UP-DU model.



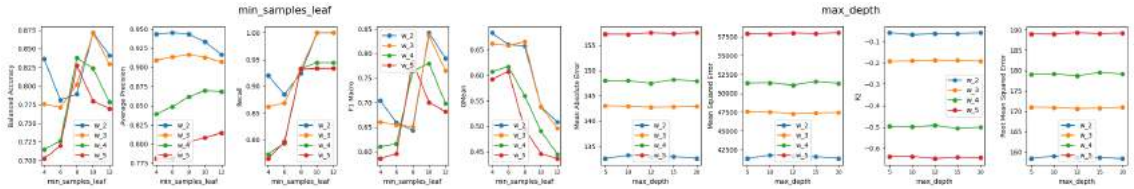
(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.102: RF hyperparameters tuning for the L5b-UP-DM model.



(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure B.103: RF hyperparameters tuning for the L5b-UP-DMU model.

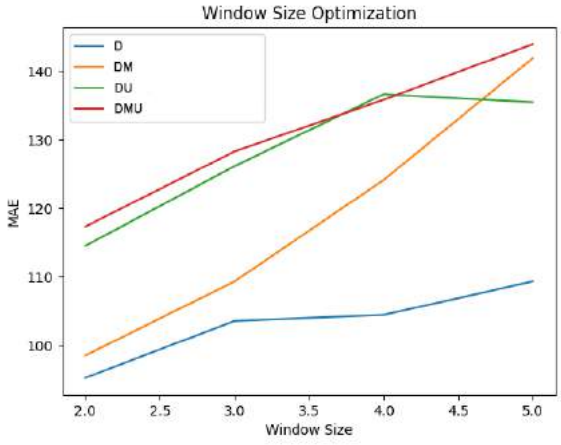
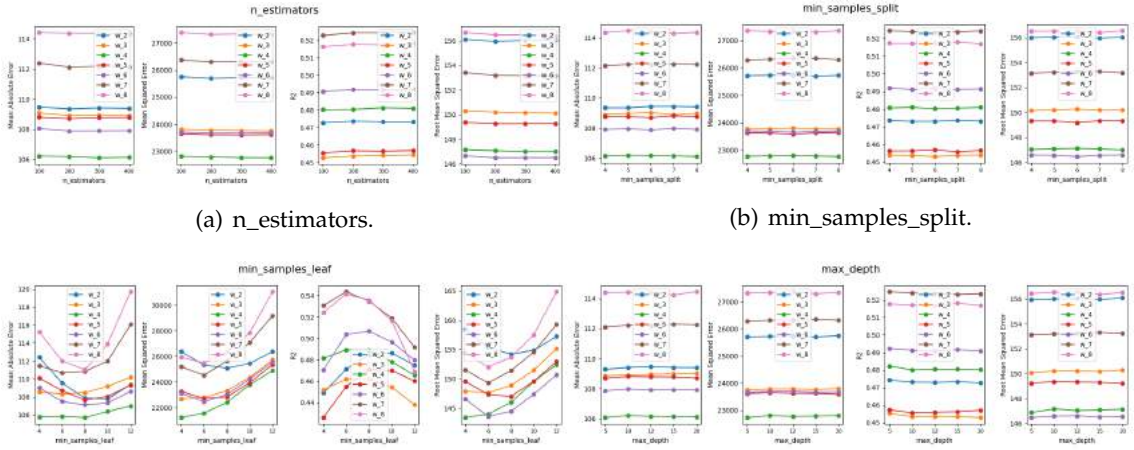


Figure B.104: L5b Caparica Window Optimization for RF Upwelling Regression models.

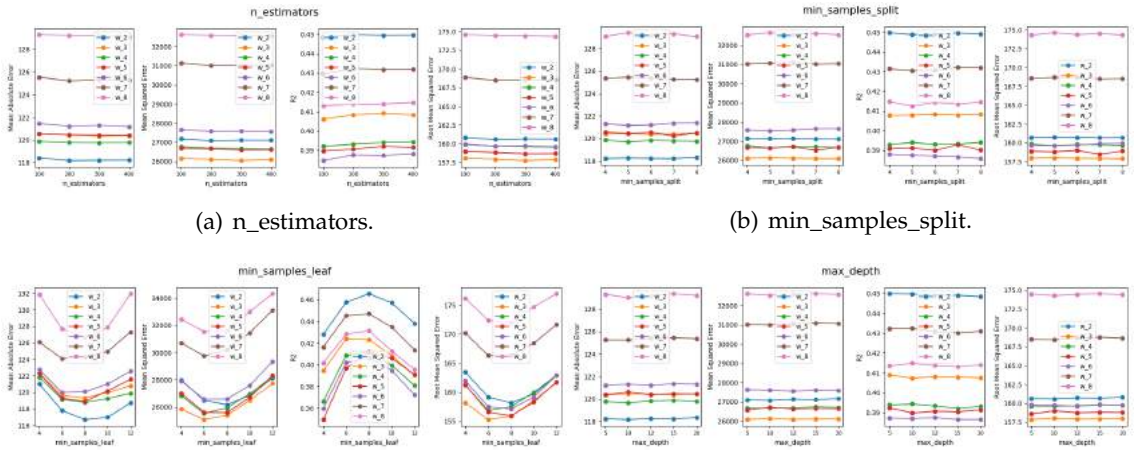
B.2.1.7 RIAV1 Triângulo RF Regression



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

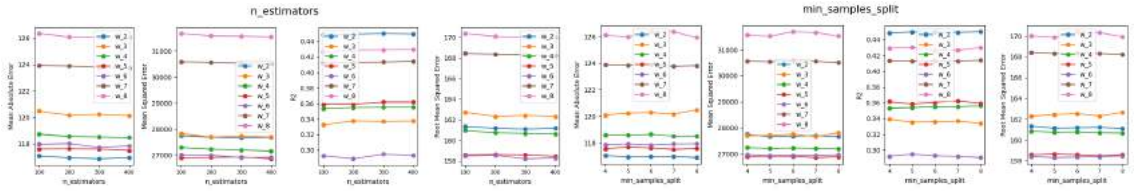
Figure B.105: RF hyperparameters tuning for the RIAV1-D model.



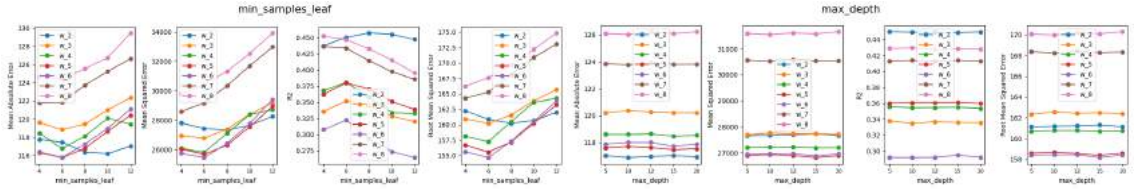
(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.106: RF hyperparameters tuning for the RIAV1-DM model.

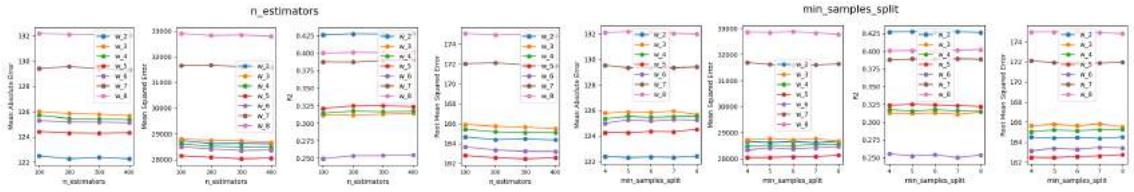


(a) n_estimators. (b) min_samples_split.

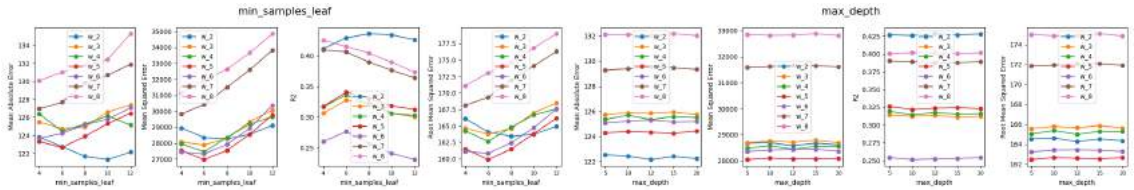


(c) min_samples_leaf. (d) max_depth.

Figure B.107: RF hyperparameters tuning for the RIAV1-DH model.



(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

Figure B.108: RF hyperparameters tuning for the RIAV1-DMH model.

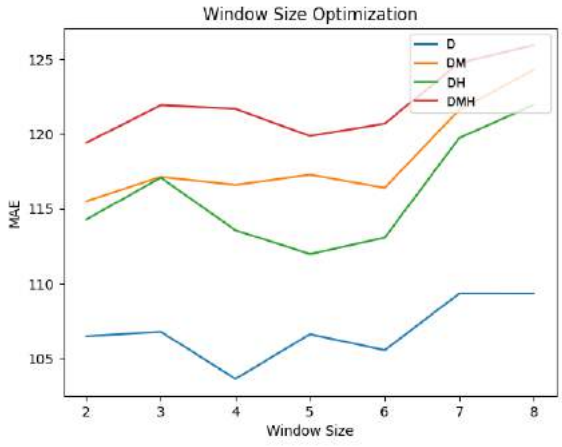
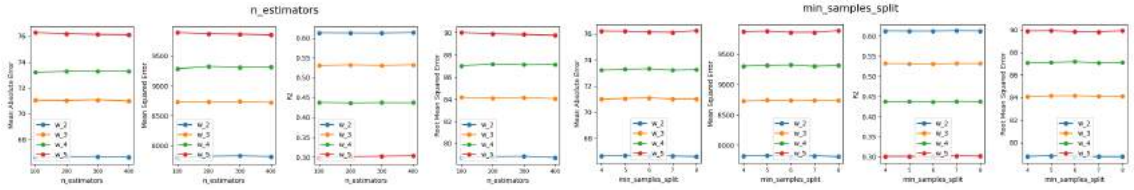
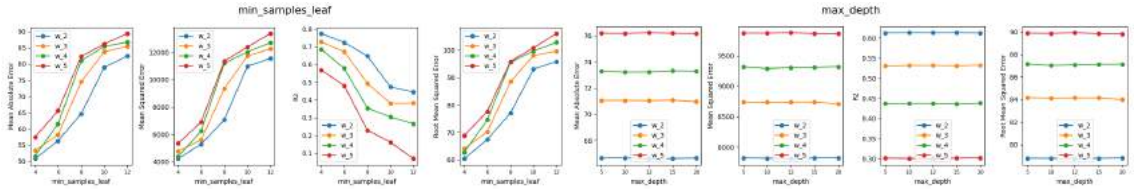


Figure B.109: RIAV1 Triângulo Window Optimization for RF Regression models.

B.2.1.8 L7c2 Porto de Mós RF Upwelling Regression

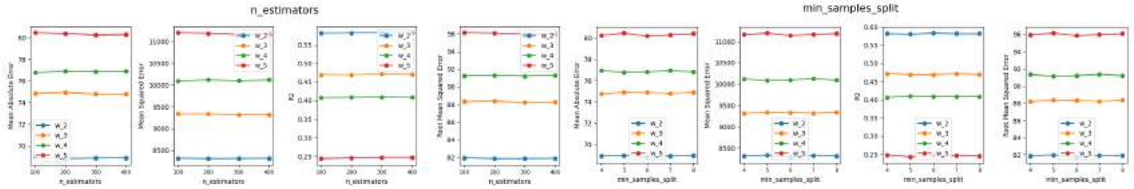


(a) n_estimators. (b) min_samples_split.

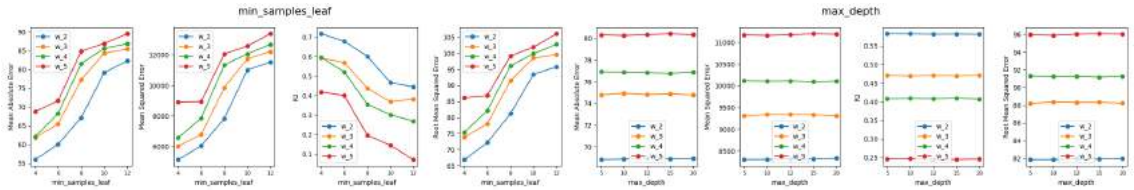


(c) min_samples_leaf. (d) max_depth.

Figure B.110: RF hyperparameters tuning for the L7c2-UP-D model.

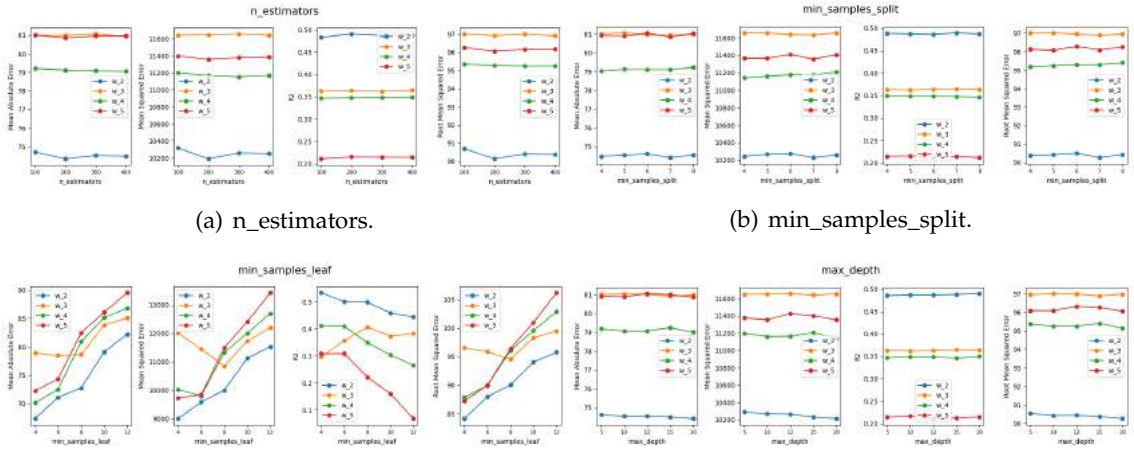


(a) n_estimators. (b) min_samples_split.



(c) min_samples_leaf. (d) max_depth.

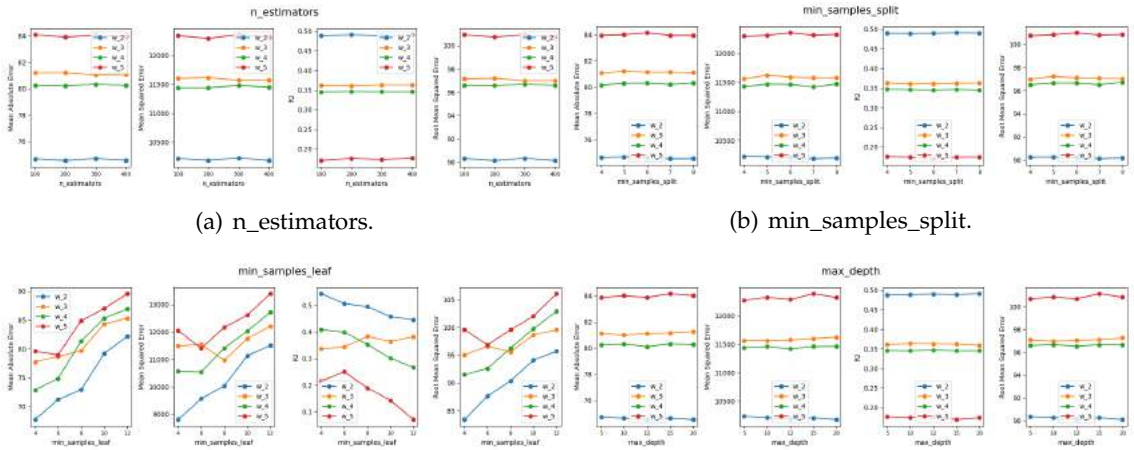
Figure B.111: RF hyperparameters tuning for the L7c2-UP-DU model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.112: RF hyperparameters tuning for the L7c2-UP-DM model.



(a) n_estimators. (b) min_samples_split.

(c) min_samples_leaf. (d) max_depth.

Figure B.113: RF hyperparameters tuning for the L7c2-UP-DMU model.

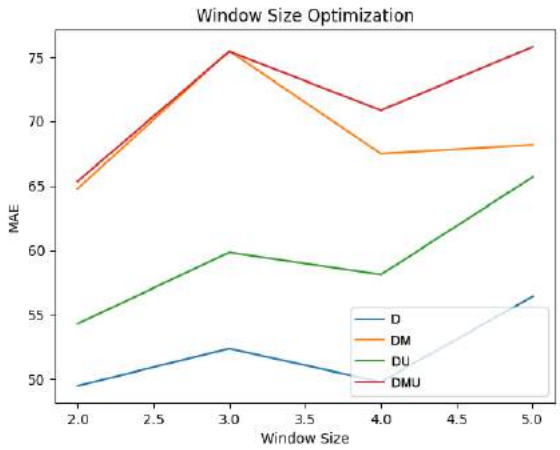
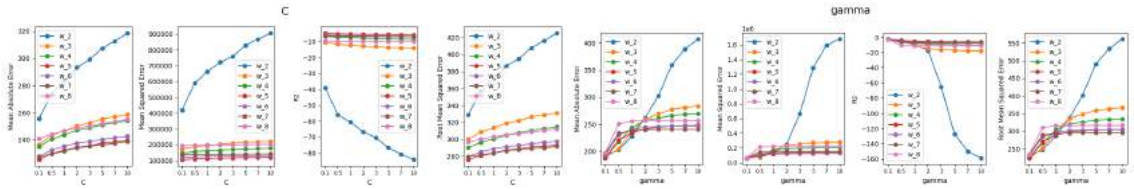


Figure B.114: L7c2 Porto de Mós Window Optimization for RF Upwelling Regression models.

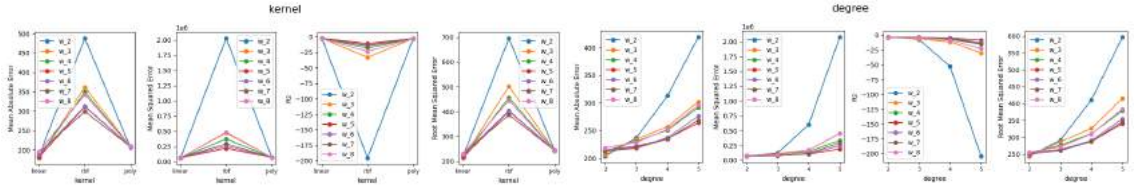
B.2.2 Support Vector Regression

B.2.2.1 L1 Carreço SVR Regression



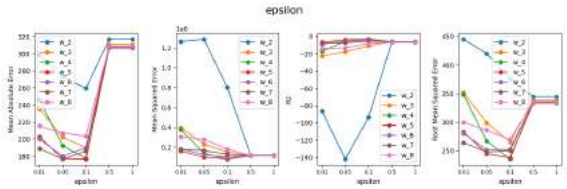
(a) C.

(b) gamma.



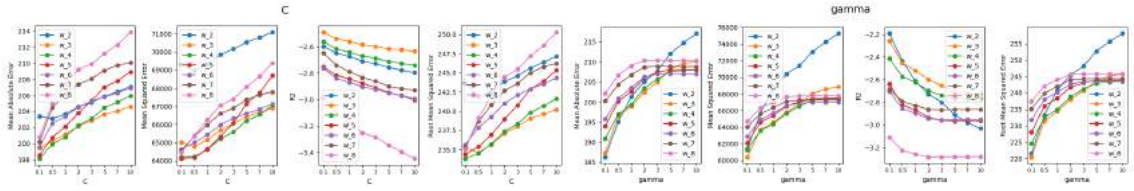
(c) kernel.

(d) degree.

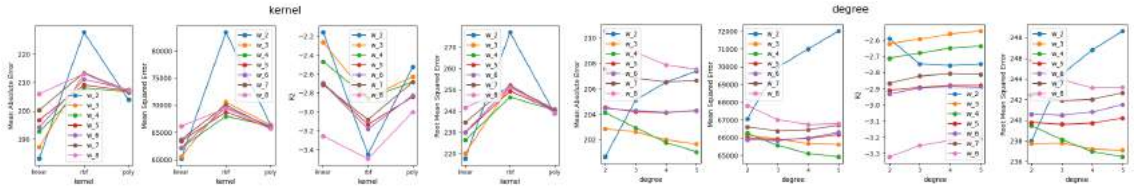


(e) epsilon.

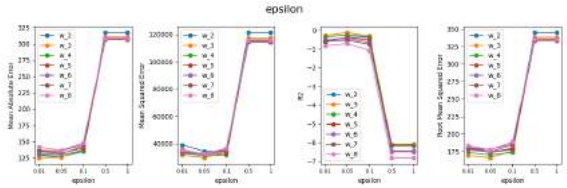
Figure B.115: SVR hyperparameters tuning for the L1-DM model.



(a) C. (b) gamma.

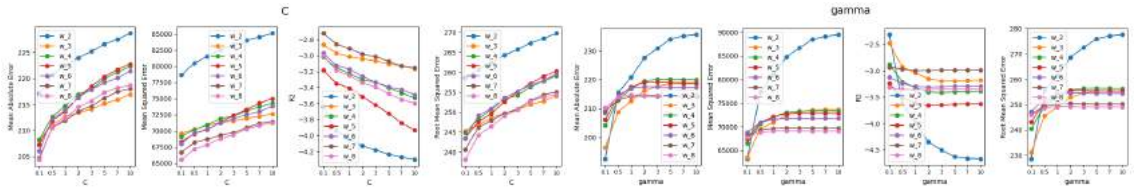


(c) kernel. (d) degree.

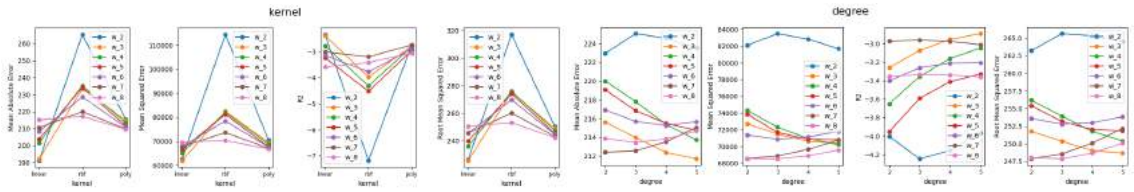


(e) epsilon.

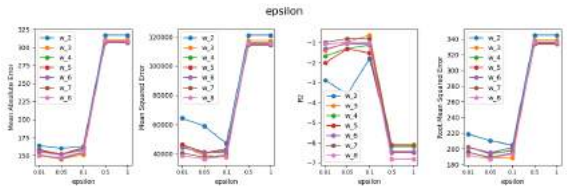
Figure B.116: SVR hyperparameters tuning for the L1-DH model.



(a) C. (b) gamma.



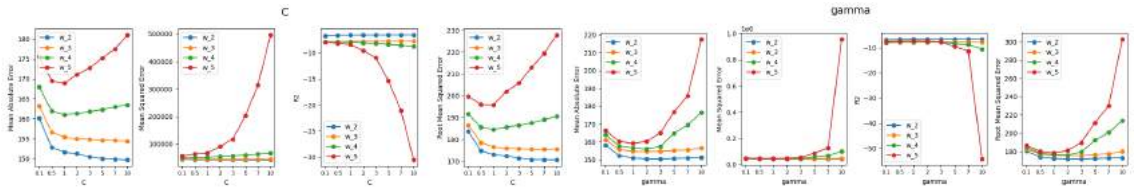
(c) kernel. (d) degree.



(e) epsilon.

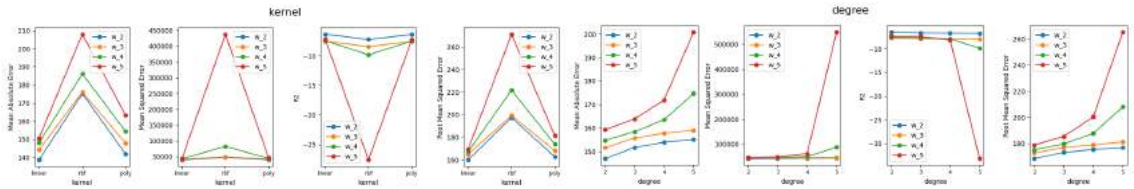
Figure B.117: SVR hyperparameters tuning for the L1-DMH model.

B.2.2.2 L1 Carreço SVR Upwelling Regression



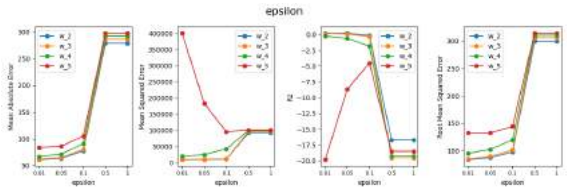
(a) C.

(b) gamma.



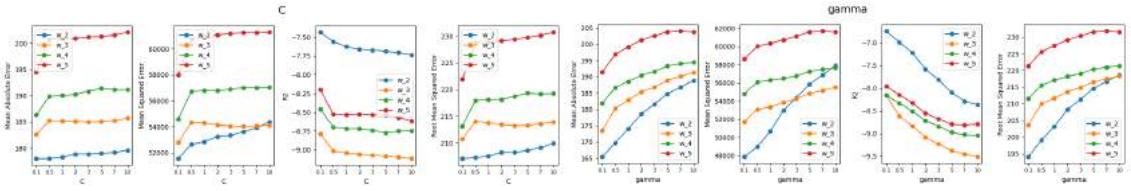
(c) kernel.

(d) degree.

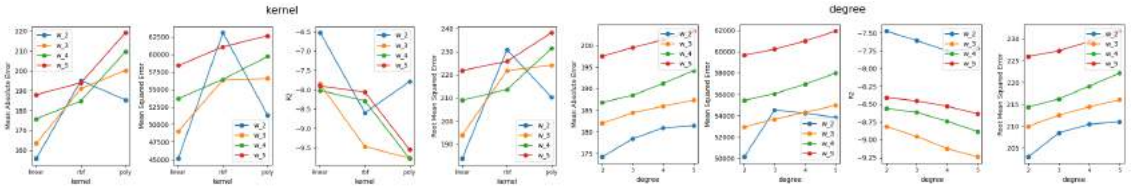


(e) epsilon.

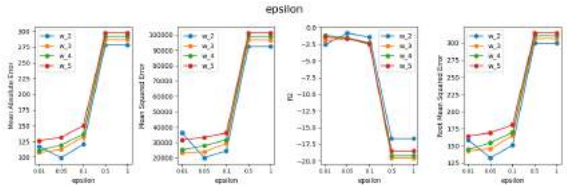
Figure B.118: SVR hyperparameters tuning for the L1-UP-DU model.



(a) C. (b) gamma.

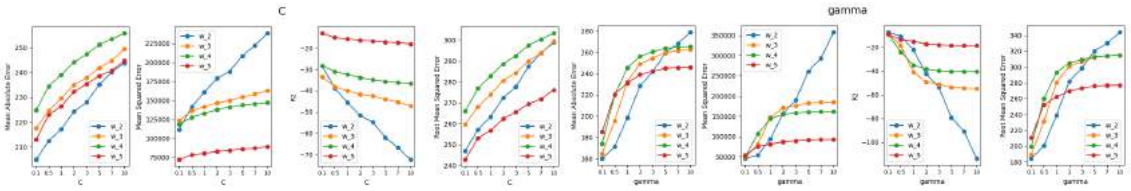


(c) kernel. (d) degree.

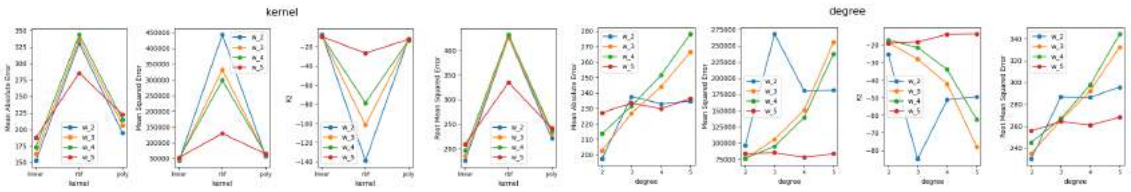


(e) epsilon.

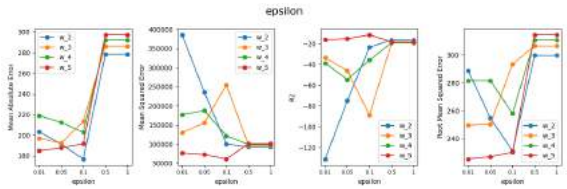
Figure B.119: SVR hyperparameters tuning for the L1-UP-DU model.



(a) C. (b) gamma.

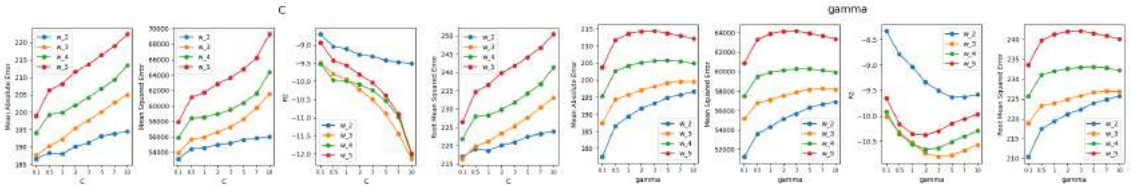


(c) kernel. (d) degree.



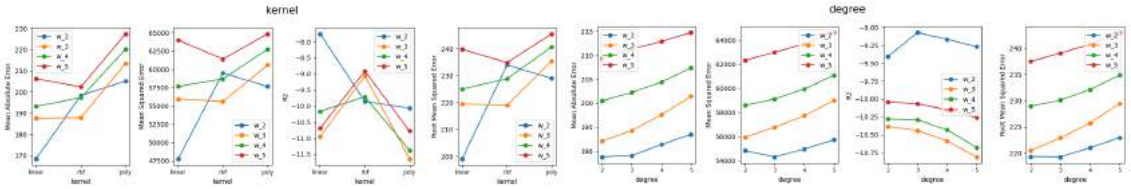
(e) epsilon.

Figure B.120: SVR hyperparameters tuning for the L1-UP-DM model.



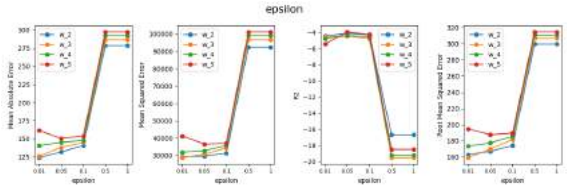
(a) C.

(b) gamma.



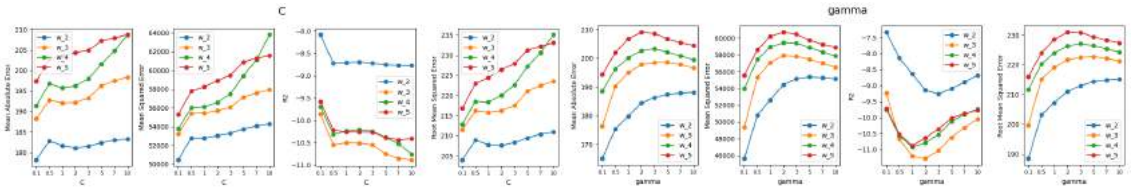
(c) kernel.

(d) degree.



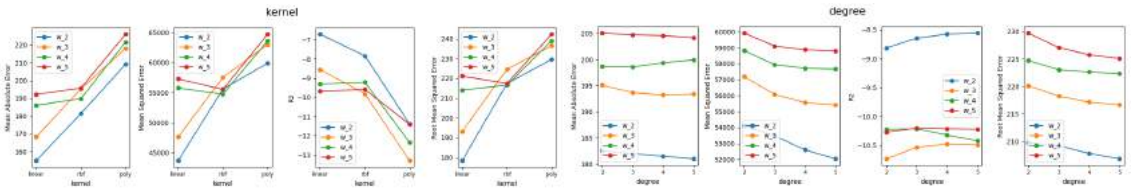
(e) epsilon.

Figure B.121: SVR hyperparameters tuning for the L1-UP-DMU model.



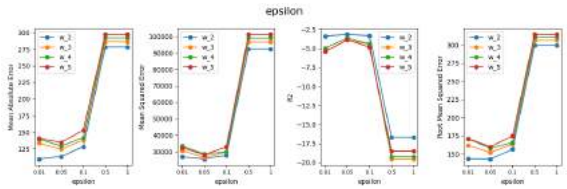
(a) C.

(b) gamma.



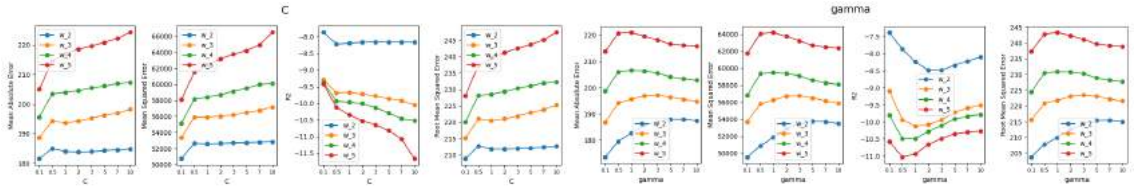
(c) kernel.

(d) degree.

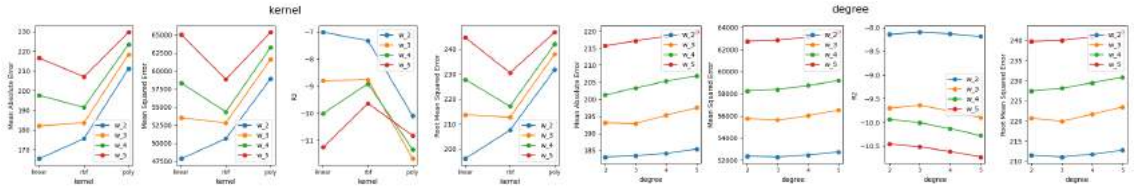


(e) epsilon.

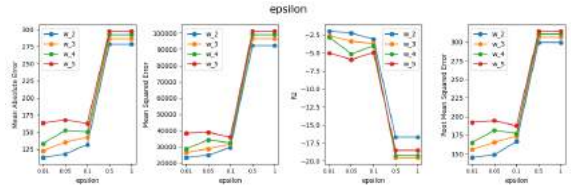
Figure B.122: SVR hyperparameters tuning for the L1-UP-DH model.



(a) C. (b) gamma.

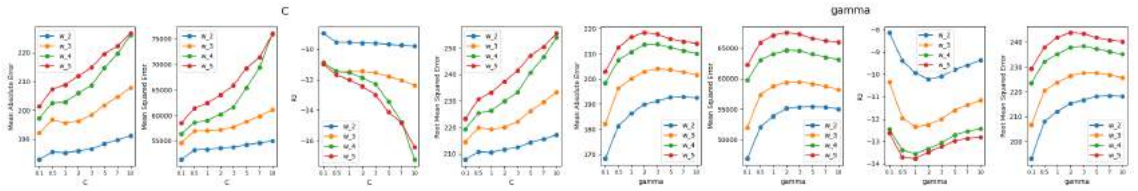


(c) kernel. (d) degree.

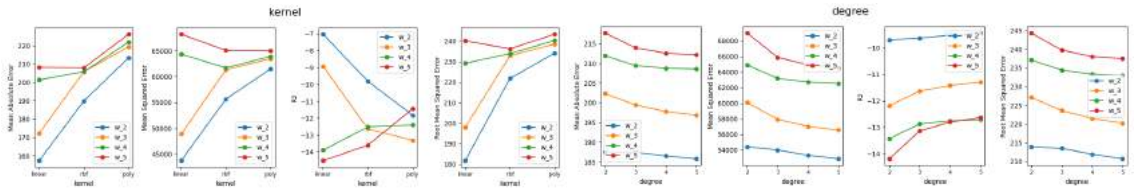


(e) epsilon.

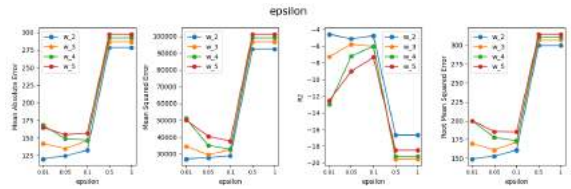
Figure B.123: SVR hyperparameters tuning for the L1-UP-DHU model.



(a) C. (b) gamma.

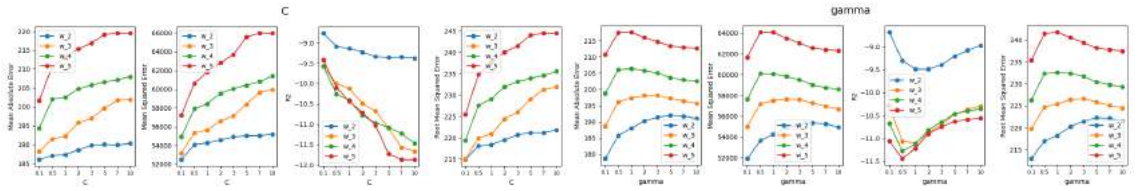


(c) kernel. (d) degree.



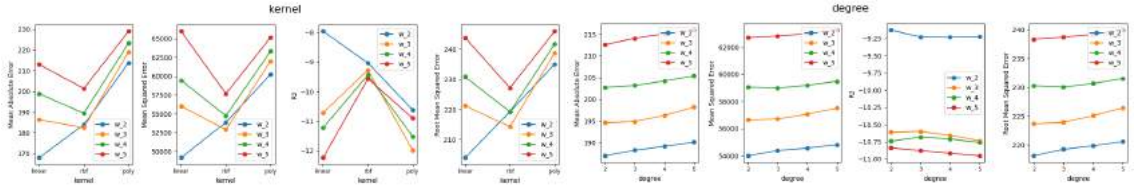
(e) epsilon.

Figure B.124: SVR hyperparameters tuning for the L1-UP-DMH model.



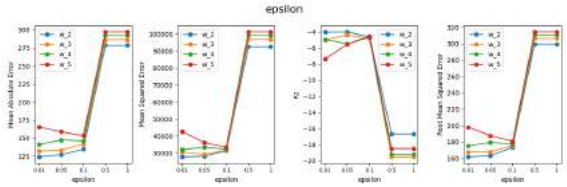
(a) C.

(b) gamma.



(c) kernel.

(d) degree.



(e) epsilon.

Figure B.125: SVR hyperparameters tuning for the L1-UP-DMHU model.

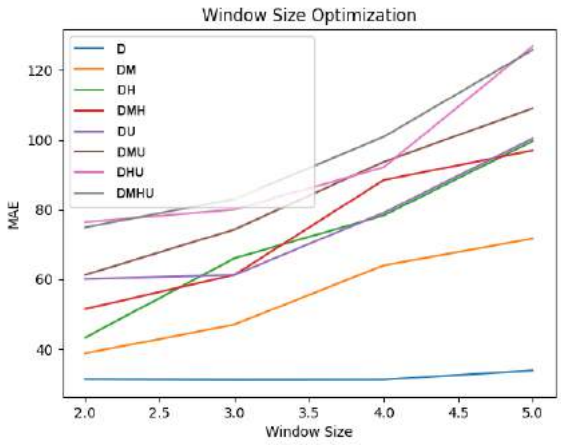
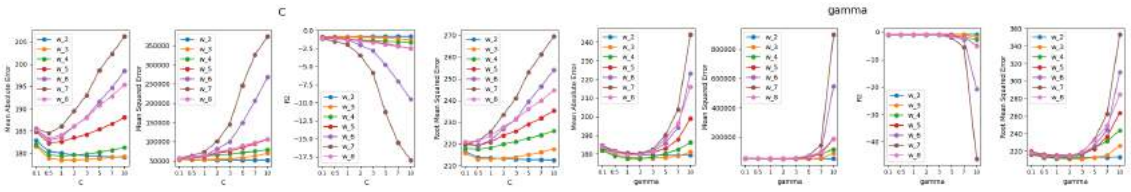


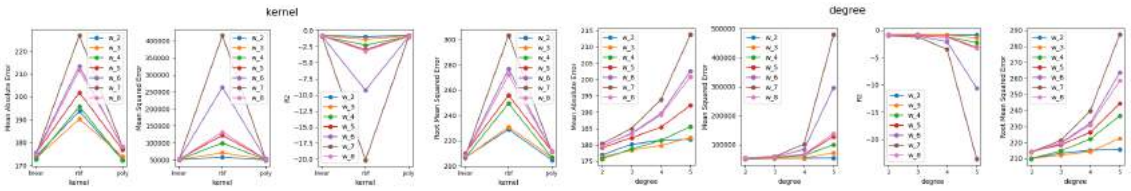
Figure B.126: L1 Carreço Window Optimization for SVR Upwelling Regression models.

B.2.2.3 L2 Leça da Palmeira SVR Regression



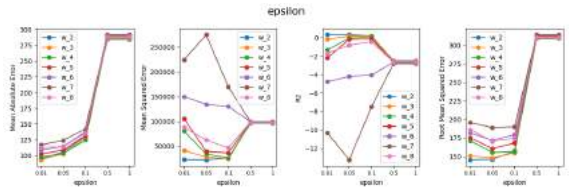
(a) C.

(b) gamma.



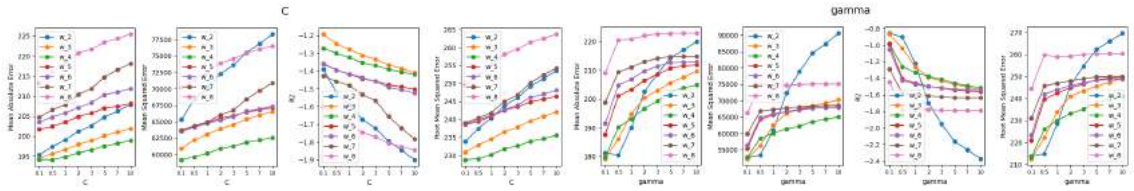
(c) kernel.

(d) degree.

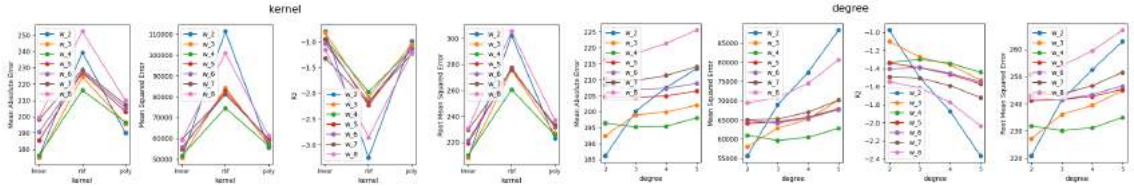


(e) epsilon.

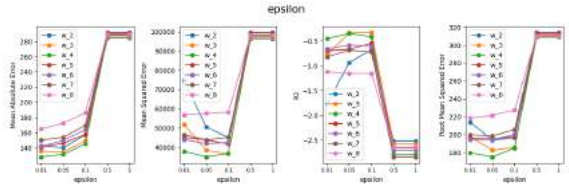
Figure B.127: SVR hyperparameters tuning for the L2-D model.



(a) C. (b) gamma.



(c) kernel. (d) degree.



(e) epsilon.

Figure B.128: SVR hyperparameters tuning for the L2-DM model.

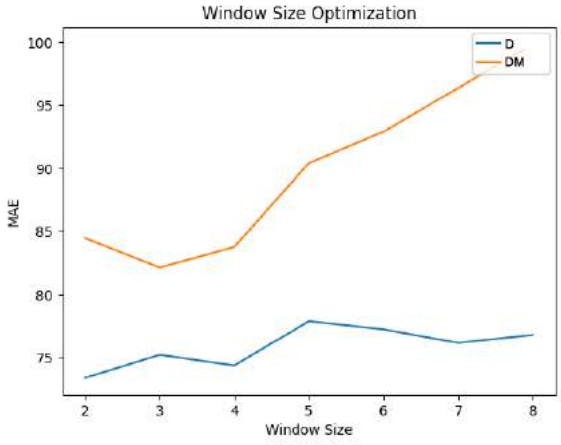
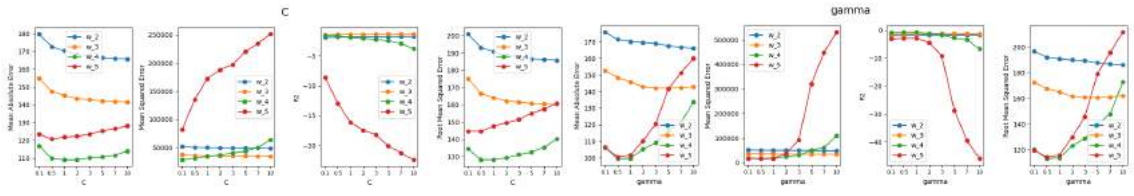


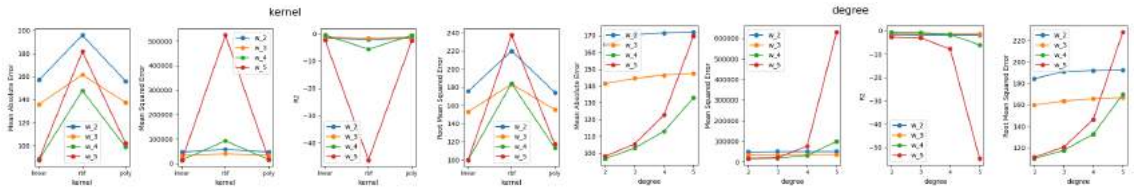
Figure B.129: L2 Leça da Palmeira Window Optimization for SVR Regression models.

B.2.2.4 L2 Leça da Palmeira SVR Upwelling Regression



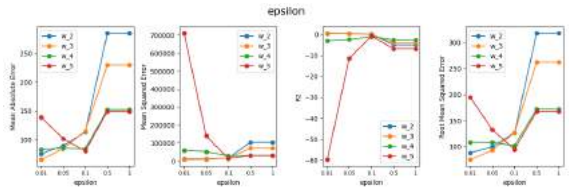
(a) C.

(b) gamma.



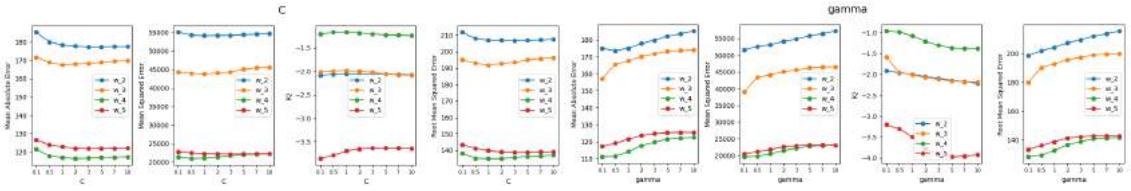
(c) kernel.

(d) degree.

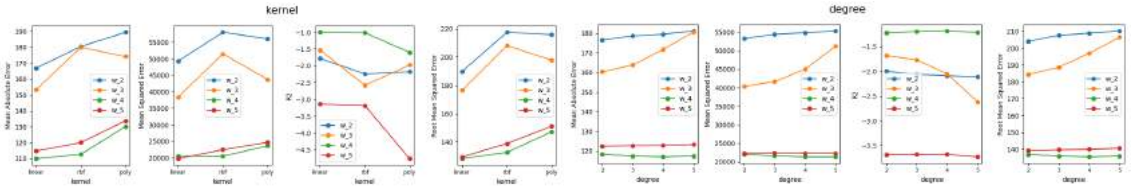


(e) epsilon.

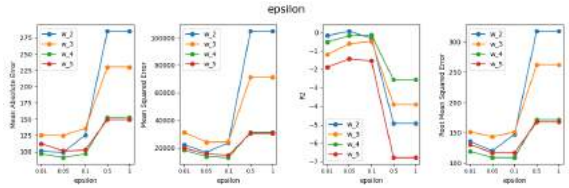
Figure B.130: SVR hyperparameters tuning for the L2-UP-D model.



(a) C. (b) gamma.

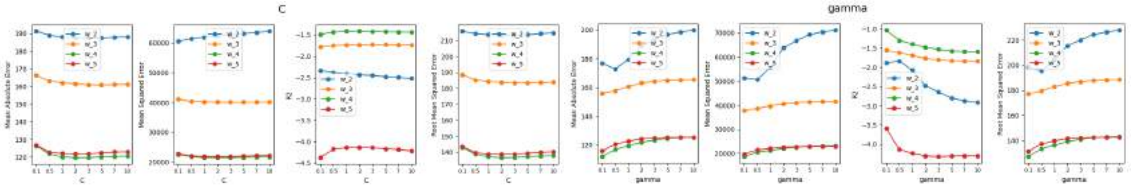


(c) kernel. (d) degree.

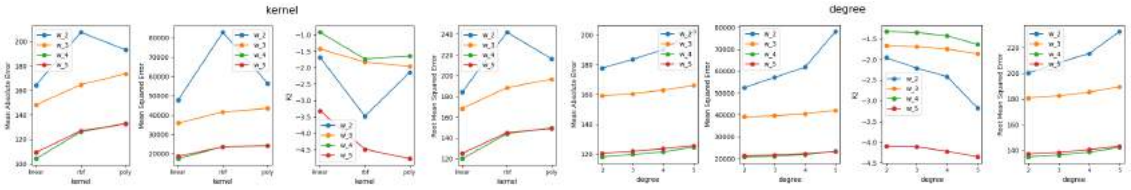


(e) epsilon.

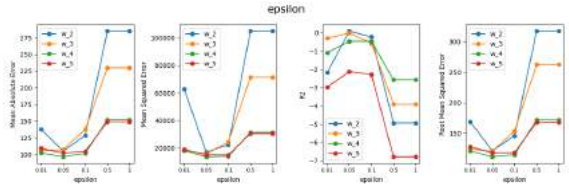
Figure B.131: SVR hyperparameters tuning for the L2-UP-DU model.



(a) C. (b) gamma.

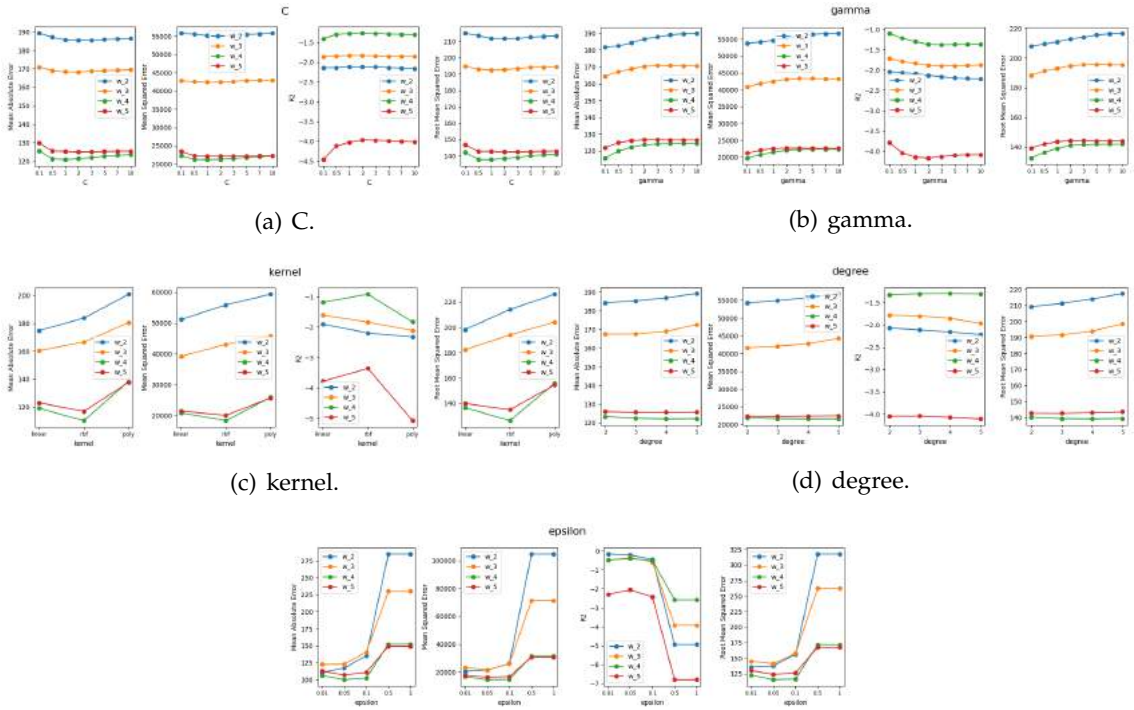


(c) kernel. (d) degree.



(e) epsilon.

Figure B.132: SVR hyperparameters tuning for the L2-UP-DM model.



(a) C.

(b) gamma.

(c) kernel.

(d) degree.

(e) epsilon.

Figure B.133: SVR hyperparameters tuning for the L2-UP-DMU model.

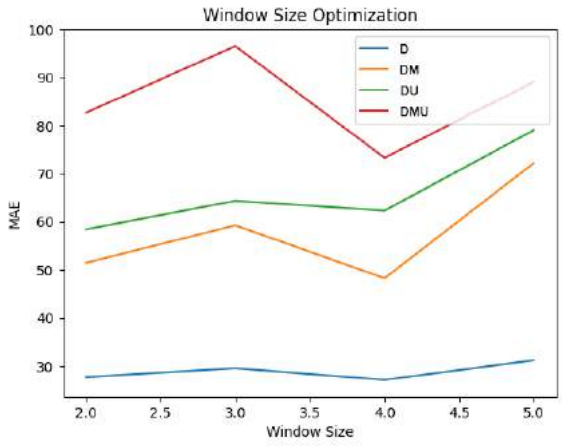
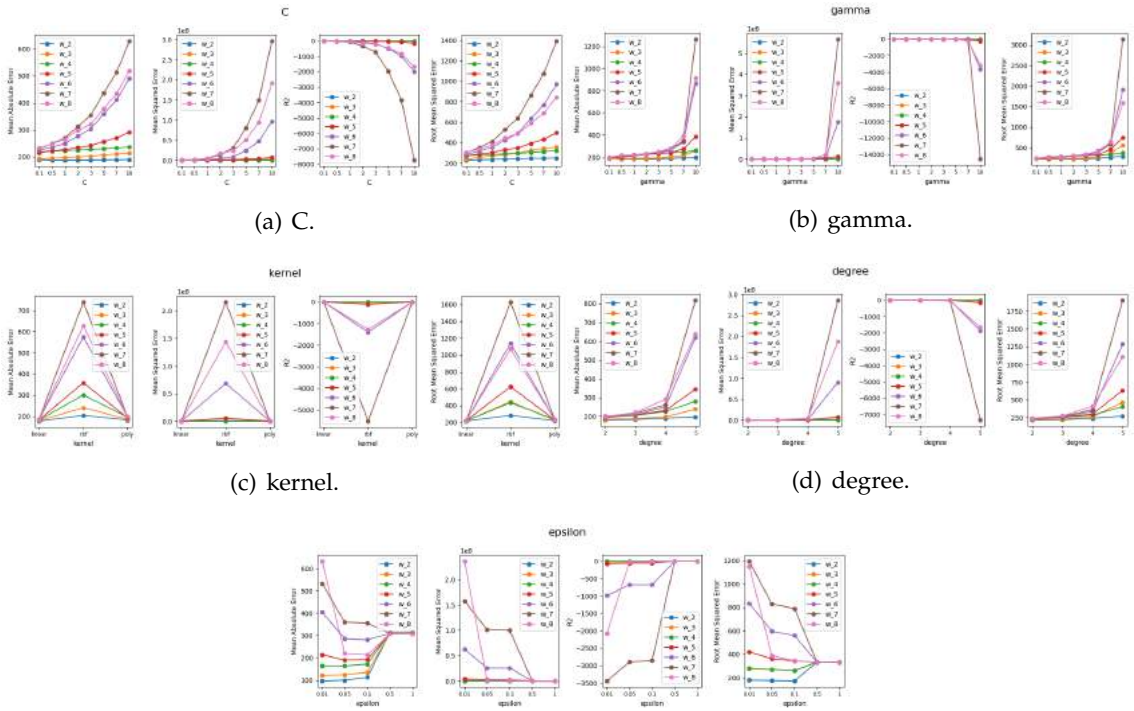


Figure B.134: L2 Leça da Palmeira Window Optimization for SVR Upwelling Regression models.

B.2.2.5 L5b Caparica SVR Regression



(a) C.

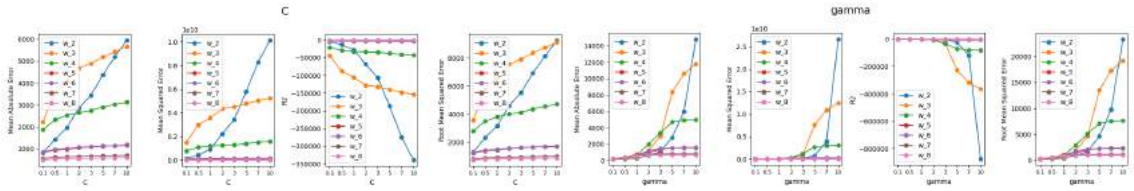
(b) gamma.

(c) kernel.

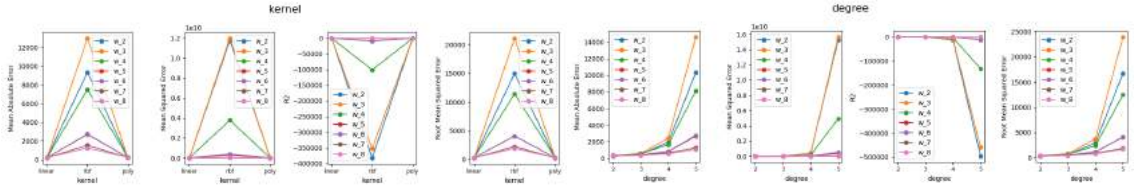
(d) degree.

(e) epsilon.

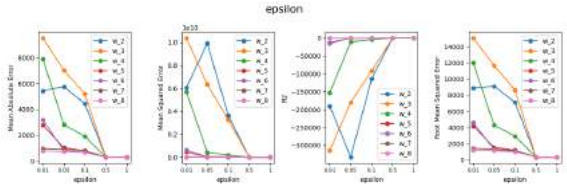
Figure B.135: SVR hyperparameters tuning for the L5b-D model.



(a) C. (b) gamma.



(c) kernel. (d) degree.



(e) epsilon.

Figure B.136: SVR hyperparameters tuning for the L5b-DM model.

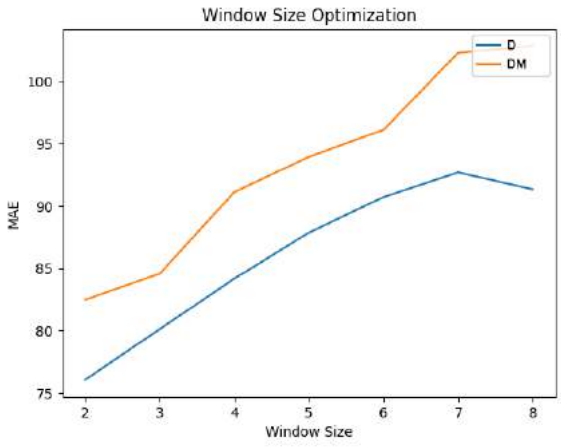
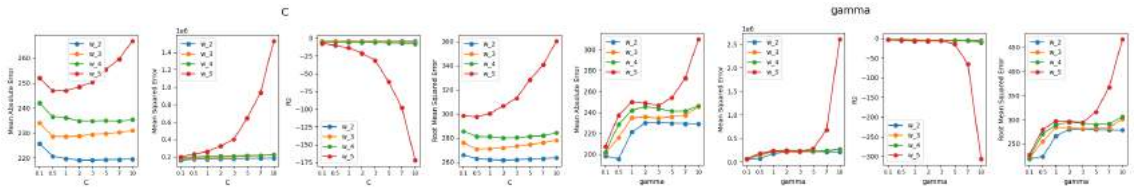


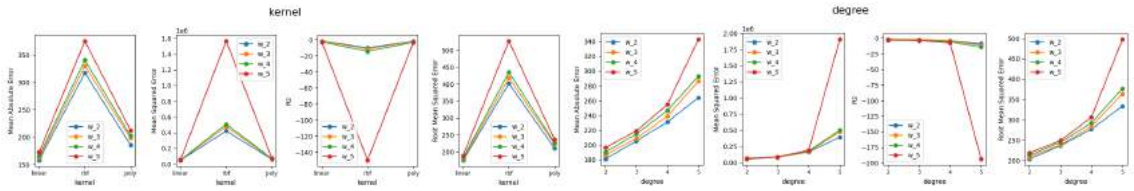
Figure B.137: L5b Caparica Window Optimization for SVR Regression models.

B.2.2.6 L5b Caparica SVR Upwelling Regression



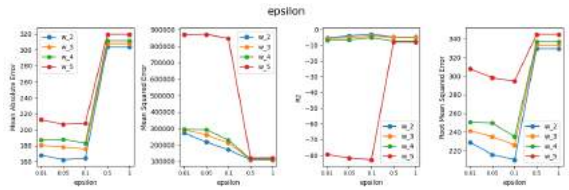
(a) C.

(b) gamma.



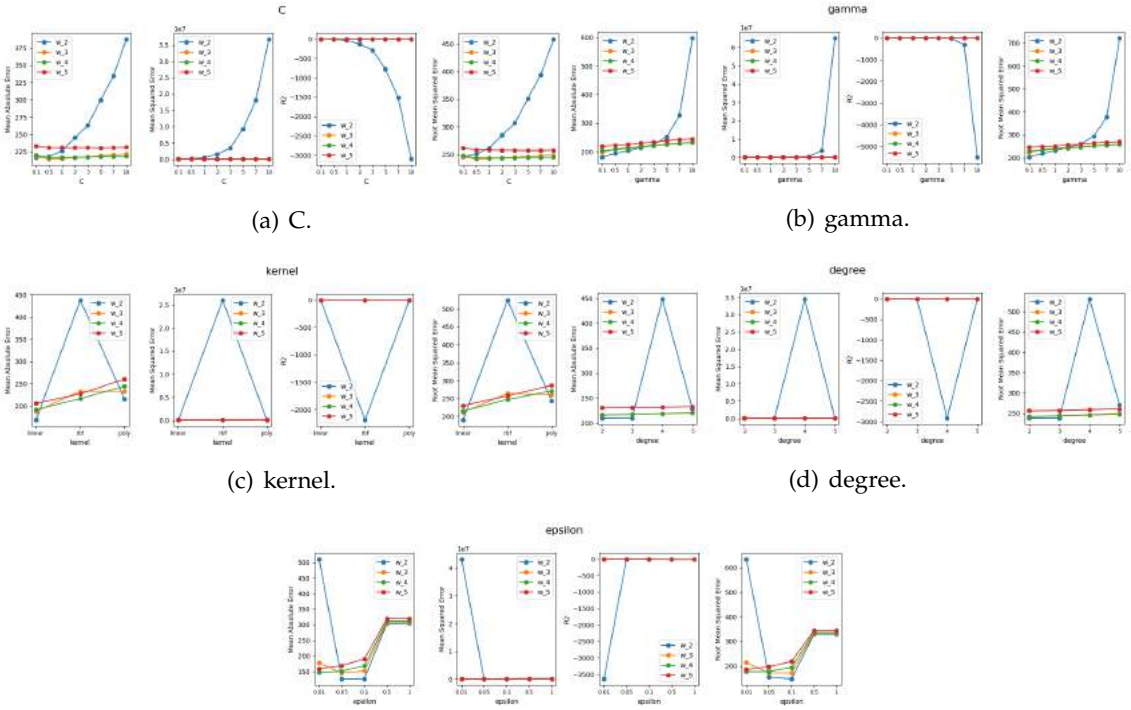
(c) kernel.

(d) degree.



(e) epsilon.

Figure B.138: SVR hyperparameters tuning for the L5b-UP-D model.



(a) C.

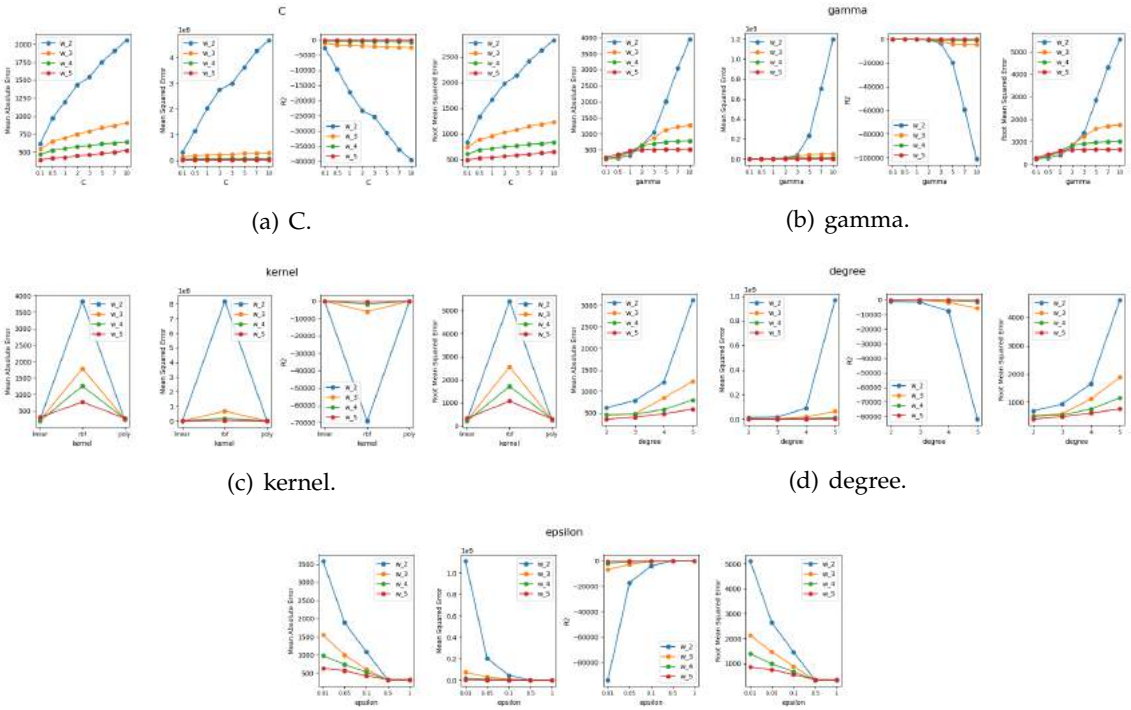
(b) gamma.

(c) kernel.

(d) degree.

(e) epsilon.

Figure B.139: SVR hyperparameters tuning for the L5b-UP-DU model.



(a) C.

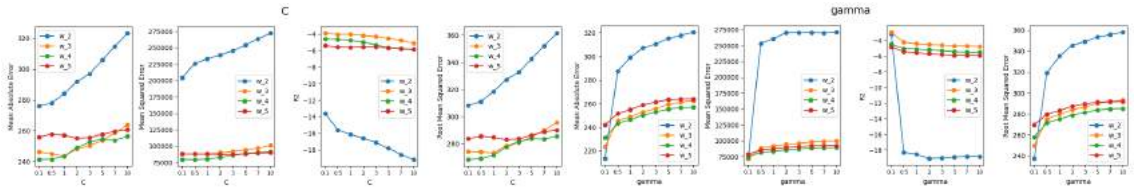
(b) gamma.

(c) kernel.

(d) degree.

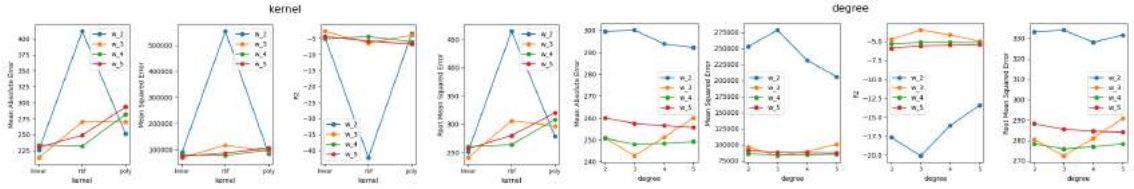
(e) epsilon.

Figure B.140: SVR hyperparameters tuning for the L5b-UP-DM model.



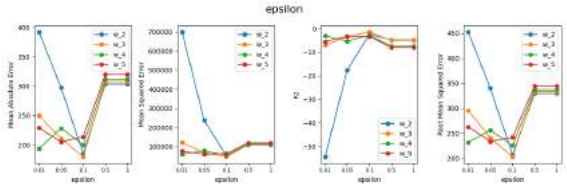
(a) C.

(b) gamma.



(c) kernel.

(d) degree.



(e) epsilon.

Figure B.141: SVR hyperparameters tuning for the L5b-UP-DMU model.

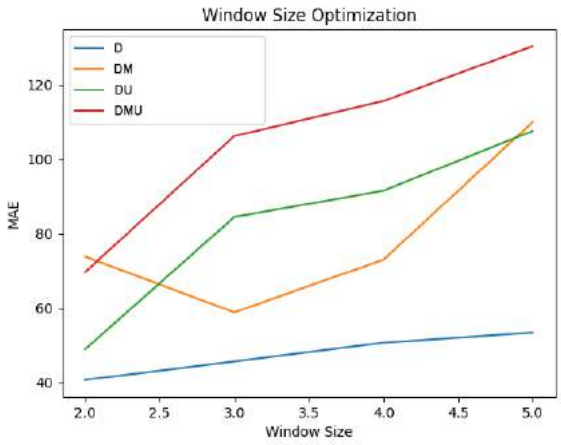
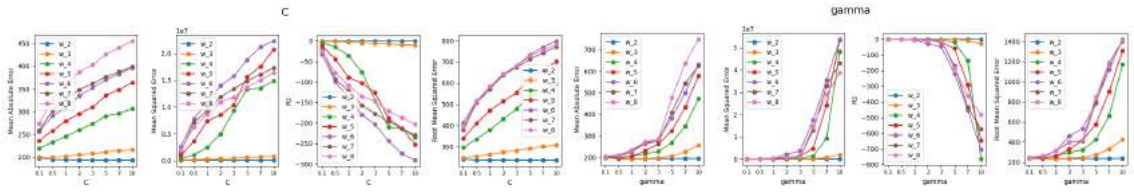


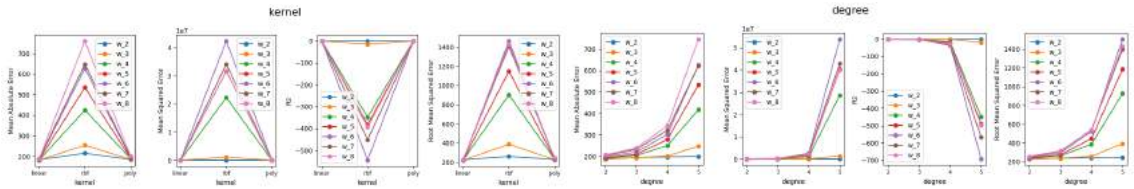
Figure B.142: L5b Caparica Window Optimization for SVR Upwelling Regression models.

B.2.2.7 RIAV1 Triângulo SVR Regression



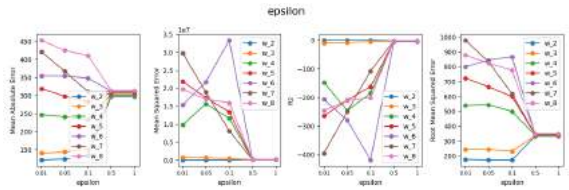
(a) C.

(b) gamma.



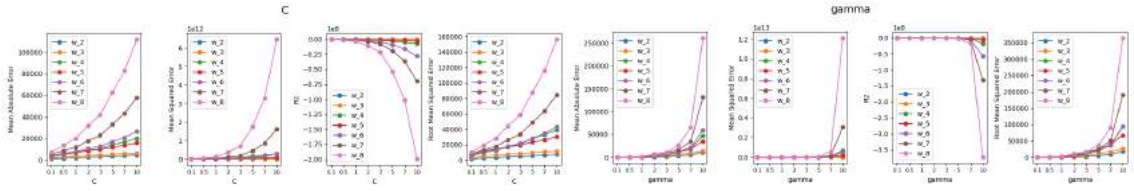
(c) kernel.

(d) degree.

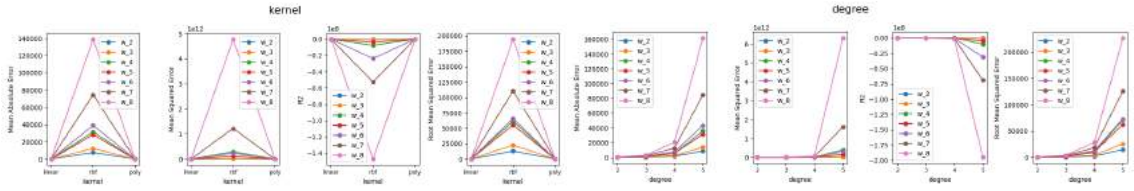


(e) epsilon.

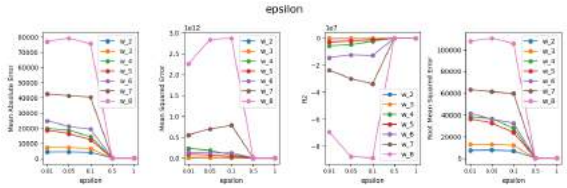
Figure B.143: SVR hyperparameters tuning for the RIAV1-D model.



(a) C. (b) gamma.

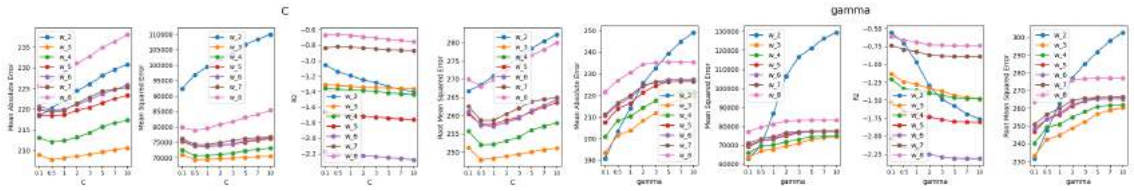


(c) kernel. (d) degree.

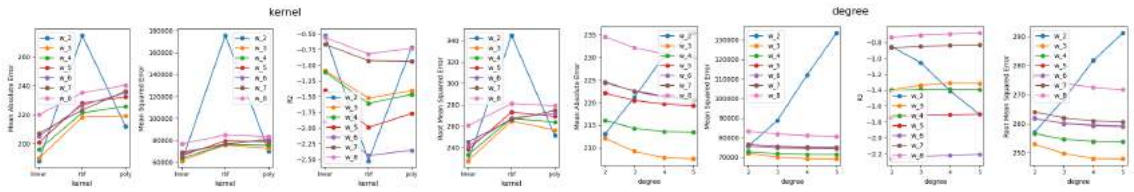


(e) epsilon.

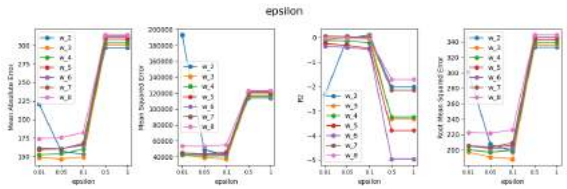
Figure B.144: SVR hyperparameters tuning for the RIAV1-DM model.



(a) C. (b) gamma.

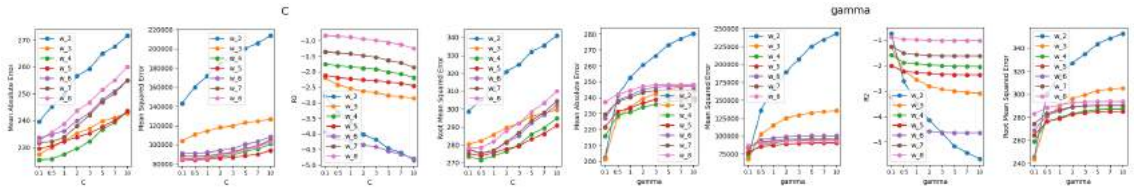


(c) kernel. (d) degree.



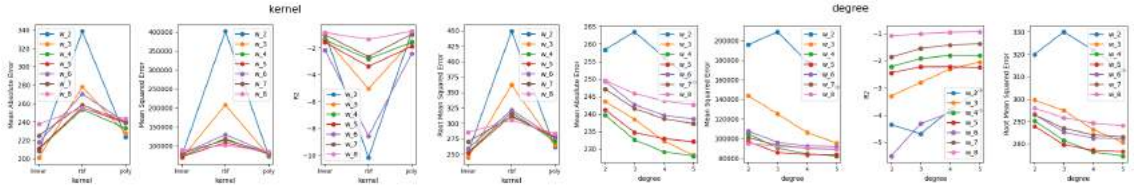
(e) epsilon.

Figure B.145: SVR hyperparameters tuning for the RIAV1-DH model.



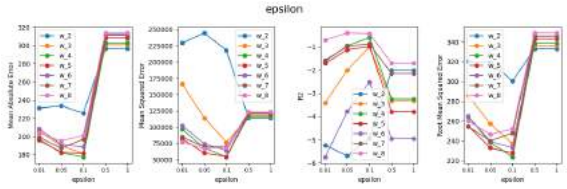
(a) C.

(b) gamma.



(c) kernel.

(d) degree.



(e) epsilon.

Figure B.146: SVR hyperparameters tuning for the RIAV1-DMH model.

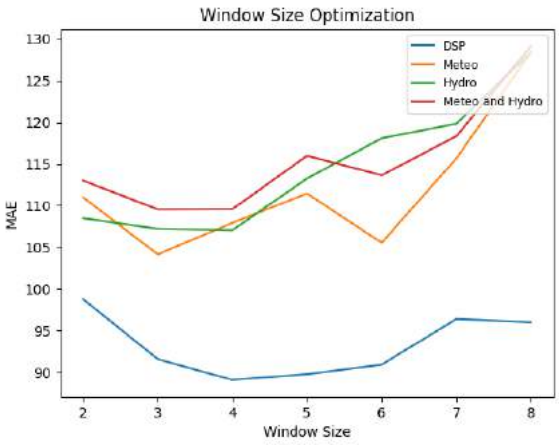
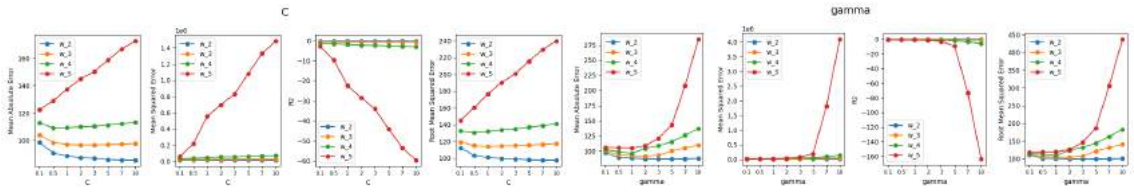


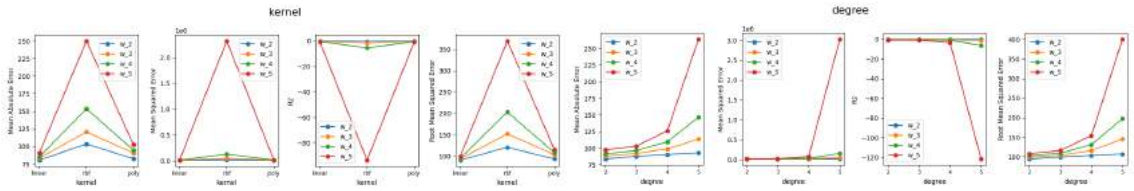
Figure B.147: RIAV1 Triângulo Window Optimization for SVR Regression models.

B.2.2.8 L7c2 Porto de Mós SVR Upwelling Regression



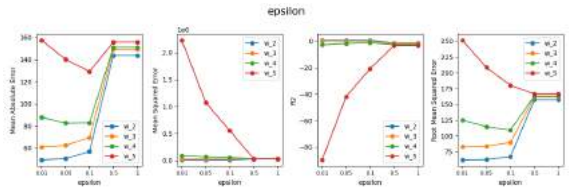
(a) C.

(b) gamma.



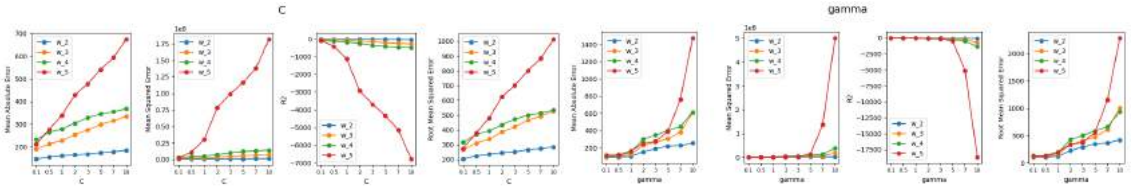
(c) kernel.

(d) degree.

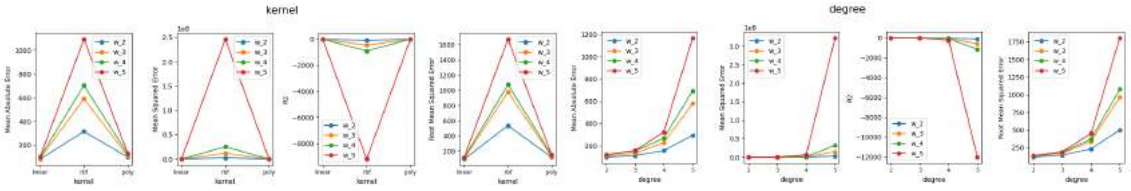


(e) epsilon.

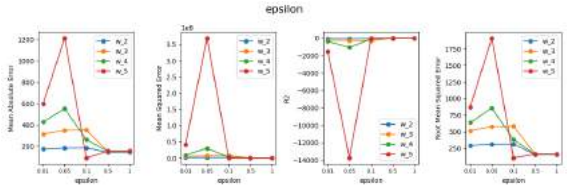
Figure B.148: SVR hyperparameters tuning for the L7c2-UP-D model.



(a) C. (b) gamma.

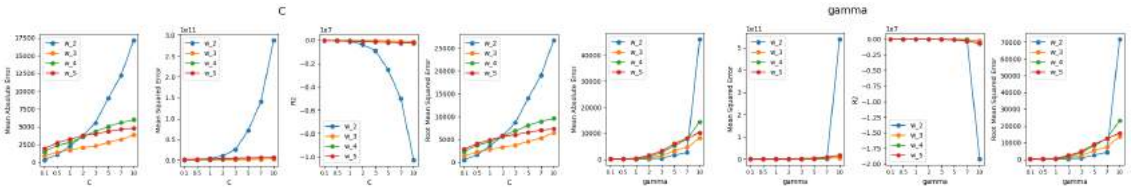


(c) kernel. (d) degree.

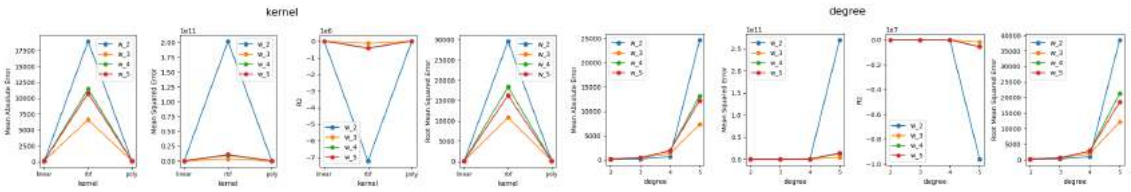


(e) epsilon.

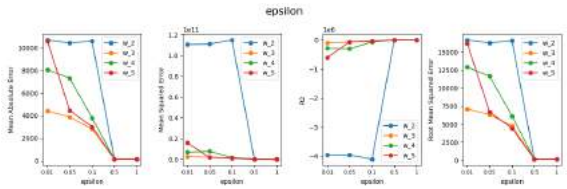
Figure B.149: SVR hyperparameters tuning for the L7c2-UP-DU model.



(a) C. (b) gamma.

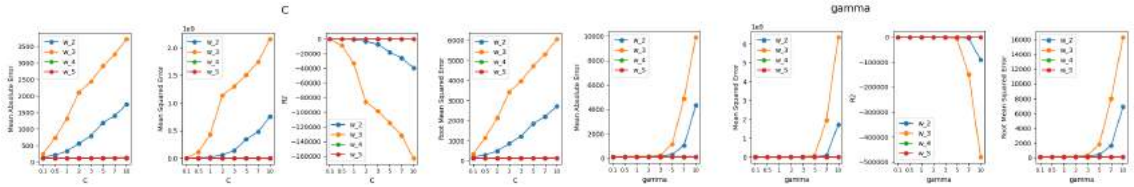


(c) kernel. (d) degree.



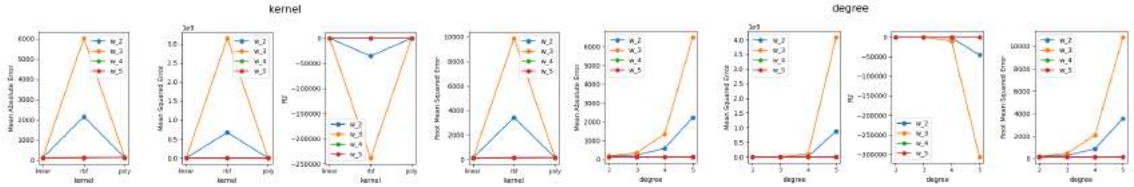
(e) epsilon.

Figure B.150: SVR hyperparameters tuning for the L7c2-UP-DM model.



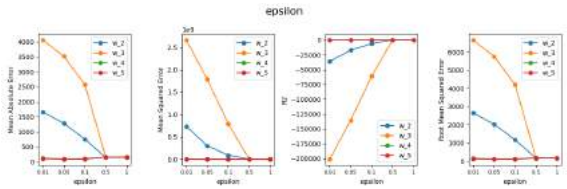
(a) C.

(b) gamma.



(c) kernel.

(d) degree.



(e) epsilon.

Figure B.151: SVR hyperparameters tuning for the L7c2-UP-DMU model.

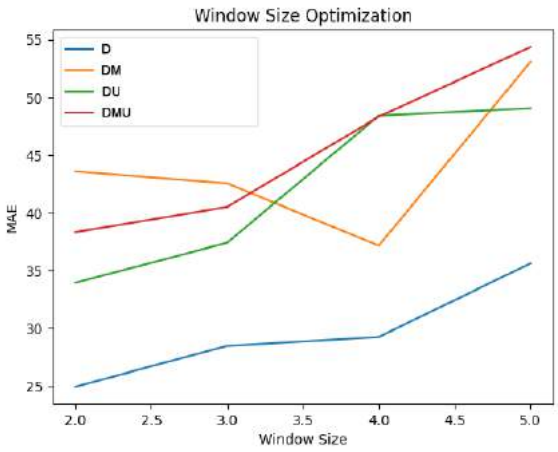


Figure B.152: L7c2 Porto de Mós Window Optimization for SVR Upwelling Regression models.

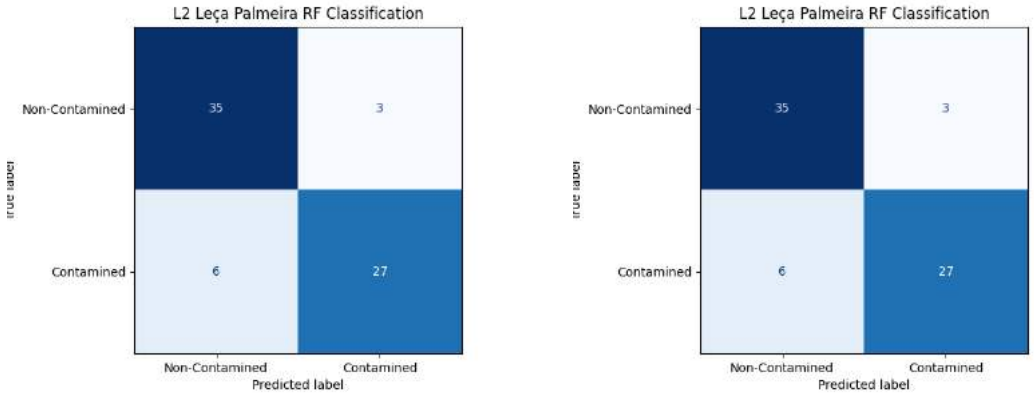
C

Appendix 3: Results

C.1 Classification

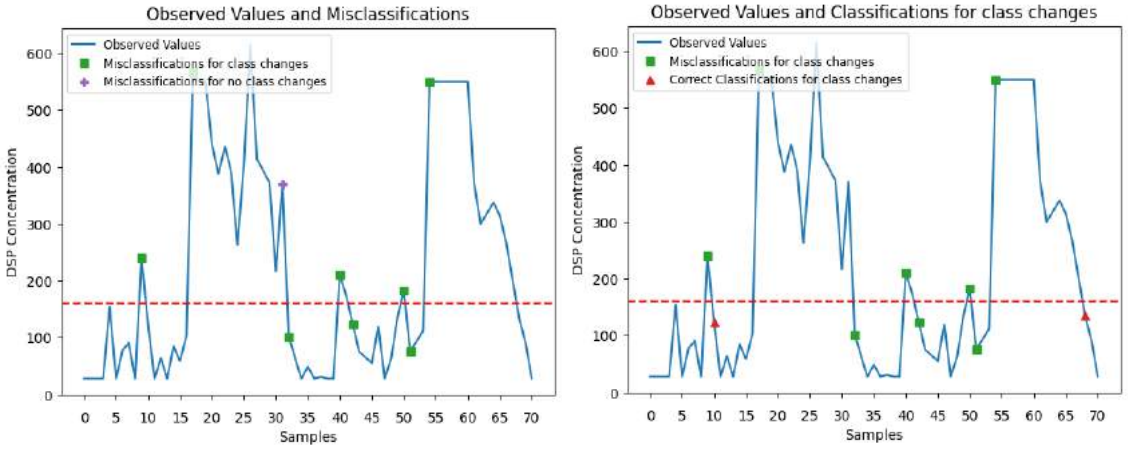
C.1.1 Random Forest

C.1.1.1 RF Classification L2 Leça da Palmeira



(a) L2-D. (b) L2-DM.

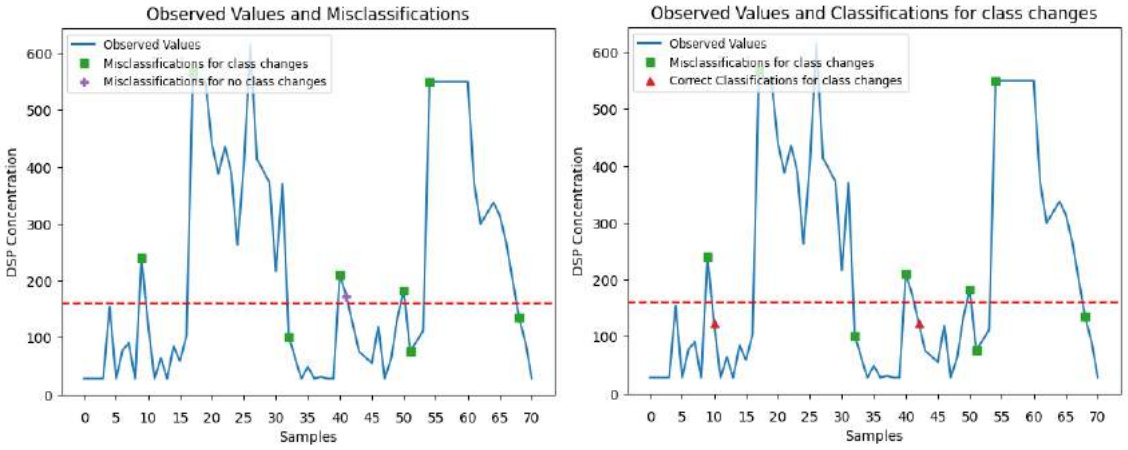
Figure C.1: RF Classification Confusion Matrices for L2 Leça da Palmeira.



(a) Misclassifications.

(b) Class Changing samples.

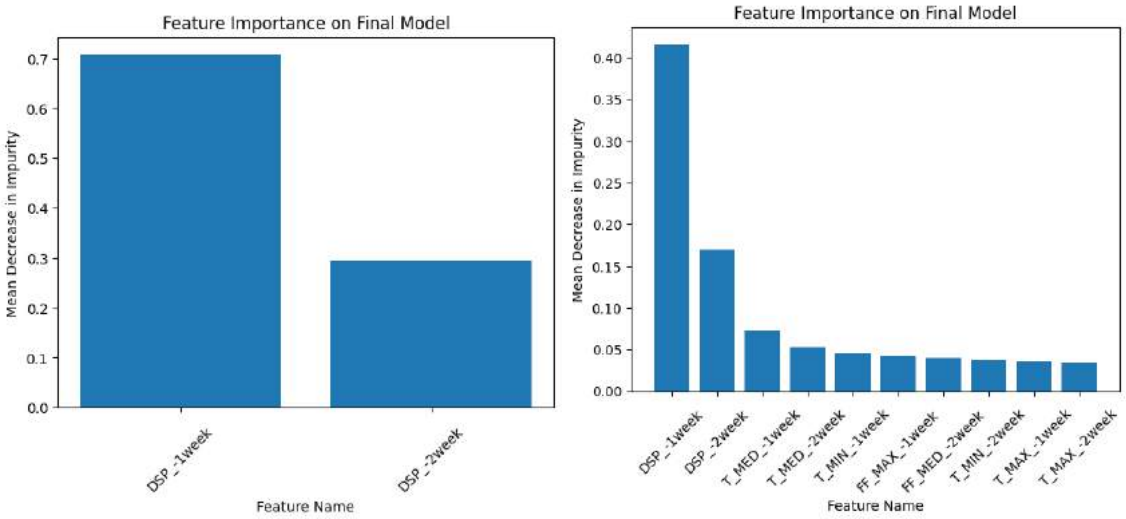
Figure C.2: Misclassifications and Class changing samples for the L2-D model.



(a) Misclassifications.

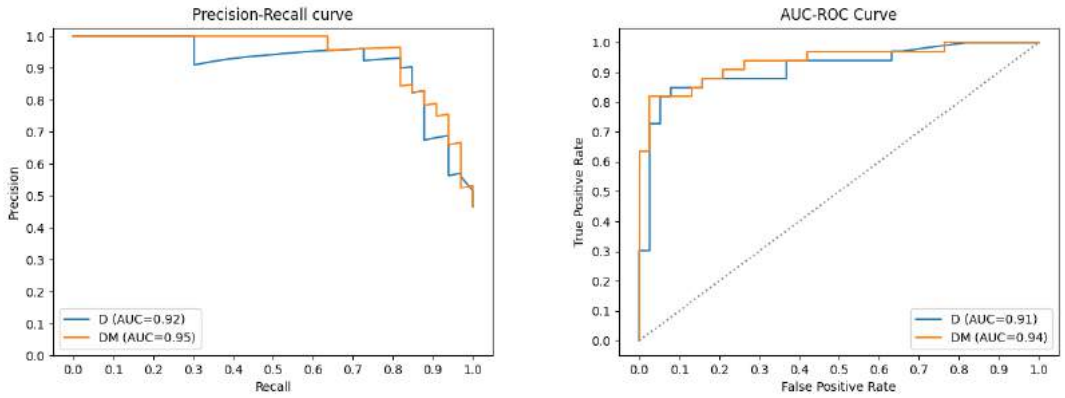
(b) Class Changing samples.

Figure C.3: Misclassifications and Class changing samples for the L2-DM model.



(a) L2-D. (b) L2-DM.

Figure C.4: L2 Leça Palmeira RF Classification feature importance.



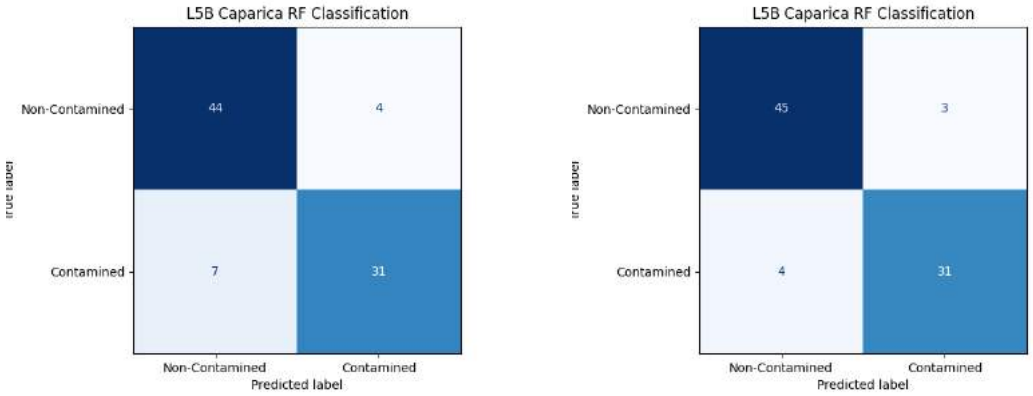
(a) PR-REC. (b) AUC-ROC.

Figure C.5: L2 Leça Palmeira RF Classification models curves.

Table C.1: RF Classification metrics for L2 Leça da Palmeira

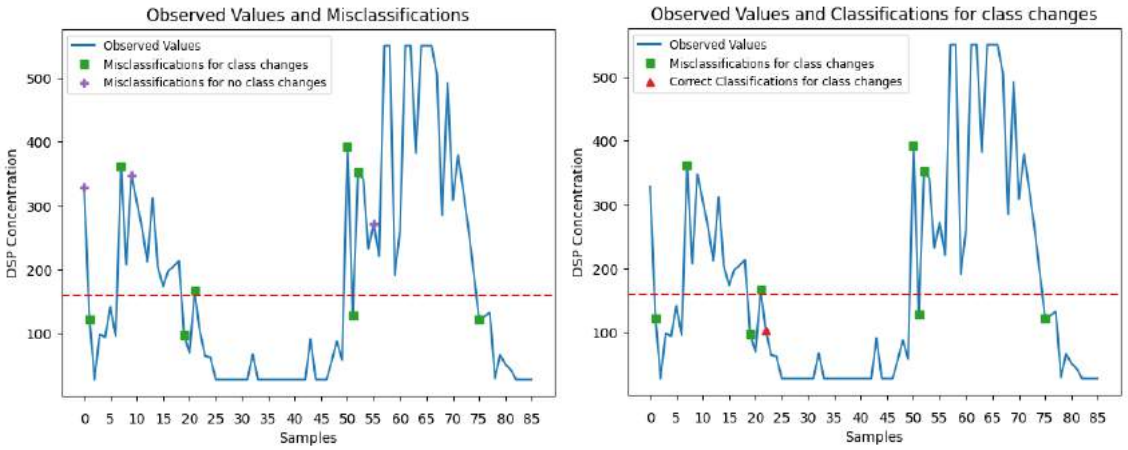
Set	Metric	Dataset	
		D	DM
Train	Balanced Accuracy	0.8478	**0.8962**
	Average Precision	0.9050	**0.9623**
	F1 Macro	0.8473	**0.8952**
	Recall	0.8306	**0.9148**
Validation	Balanced Accuracy	**0.8145**	0.7644
	Average Precision	**0.8329**	0.7778
	F1 Macro	**0.8162**	0.7689
	Recall	**0.7866**	0.7308
Test	Balanced Accuracy	**0.8696**	**0.8696**
	Average Precision	0.9154	**0.9450**
	F1 Macro	**0.8716**	**0.8716**
	Recall	**0.8182**	**0.8182**

C.1.1.2 RF Classification L5b Caparica



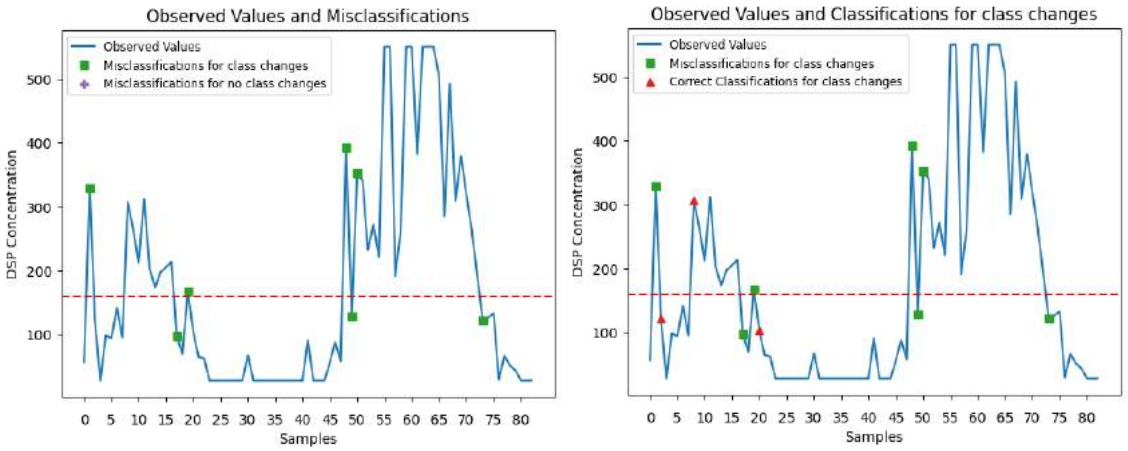
(a) L5b-D. (b) L5b-DM.

Figure C.6: RF Classification Confusion Matrices for L5b Caparica.



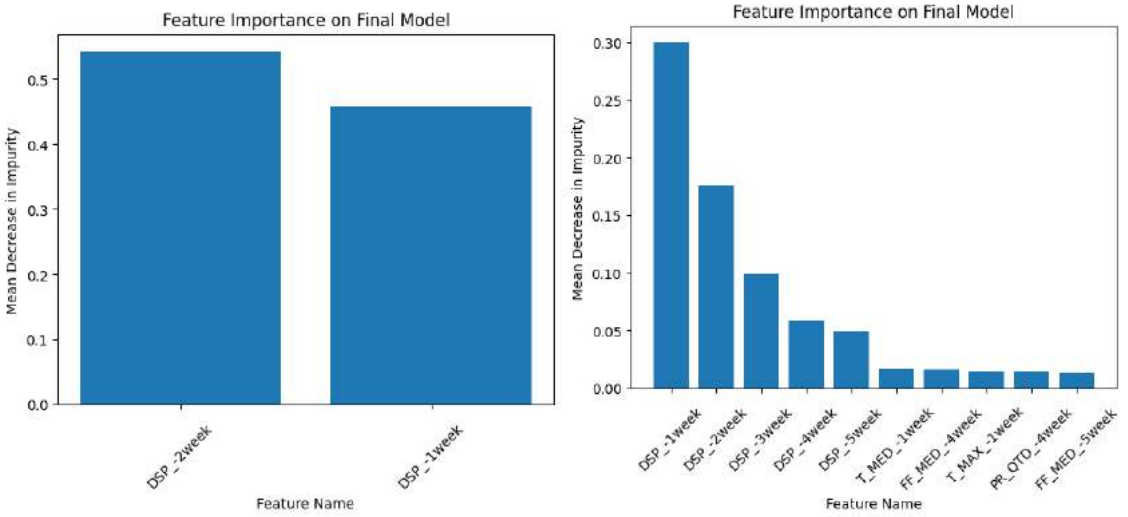
(a) Misclassifications. (b) Class Changing samples.

Figure C.7: Misclassifications and Class changing samples for the L5b-D model.



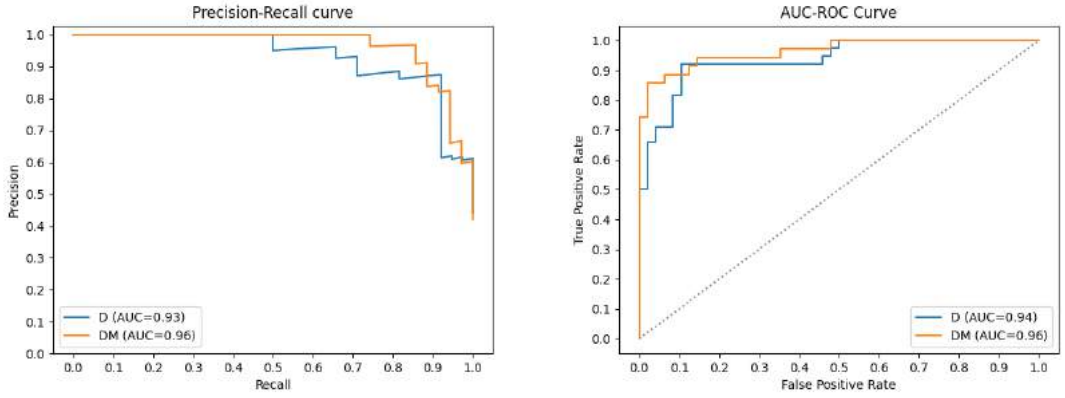
(a) Misclassifications. (b) Class Changing samples.

Figure C.8: Misclassifications and Class changing samples for the L5b-DM model.



(a) L5b-D. (b) L5b-DM.

Figure C.9: L5b Caparica RF Classification feature importance.



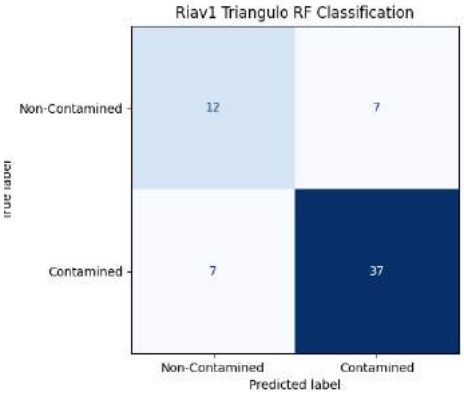
(a) PR-REC. (b) ROC-AUC.

Figure C.10: L5b Caparica RF Classification models curves.

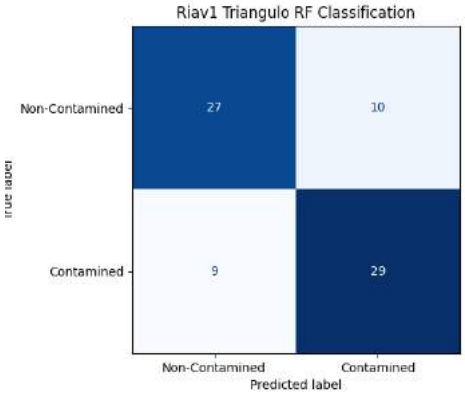
Table C.2: RF Classification metrics for L5b-Caparica

Set	Metric	Dataset	
		D	DM
Train	Balanced Accuracy	0.9125	**0.9231**
	Average Precision	0.9610	**0.9817**
	F1 Macro	0.9155	**0.9263**
	Recall	0.8789	**0.8951**
Validation	Balanced Accuracy	**0.8304**	0.8117
	Average Precision	**0.8570**	0.8193
	F1 Macro	**0.8309**	0.8183
	Recall	**0.7852**	0.7358
Test	Balanced Accuracy	0.8662	**0.9116**
	Average Precision	0.9330	**0.9632**
	F1 Macro	0.8691	**0.9132**
	Recall	0.8158	**0.8857**

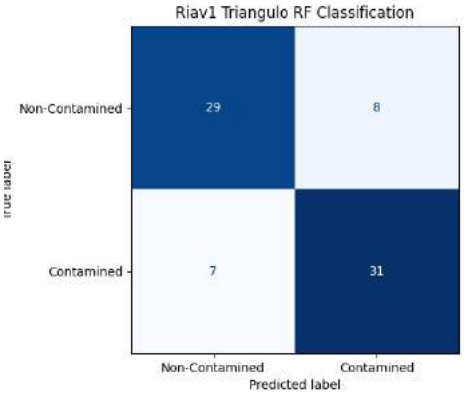
C.1.1.3 RF Classification RIAV1 Triângulo



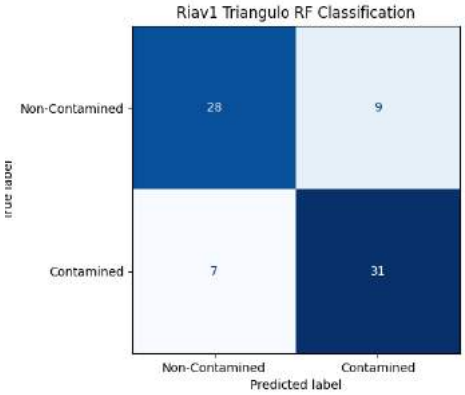
(a) RIAV1-D.



(b) RIAV1-DM.

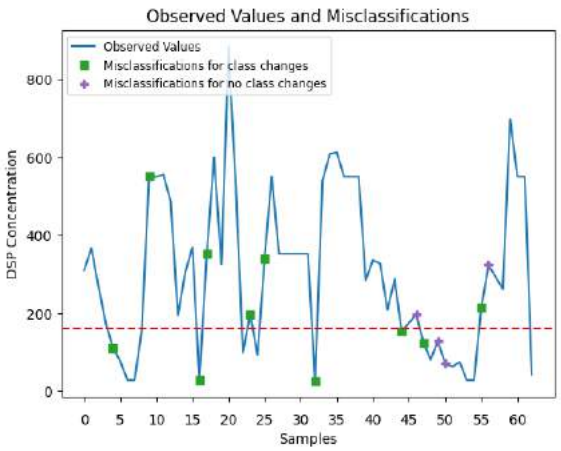


(c) RIAV1-DH.

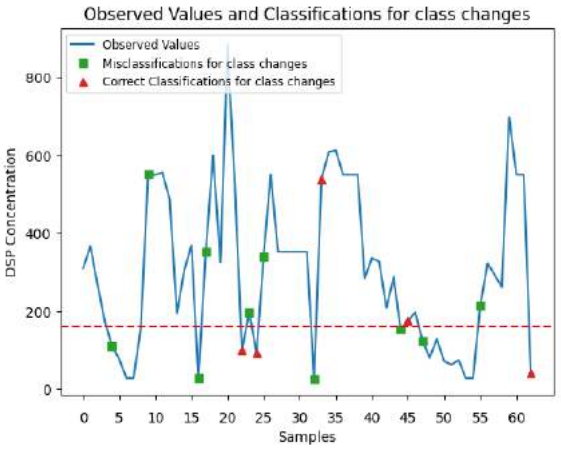


(d) RIAV1-DHM.

Figure C.11: RF Classification Confusion Matrices for RIAV1 Triângulo.

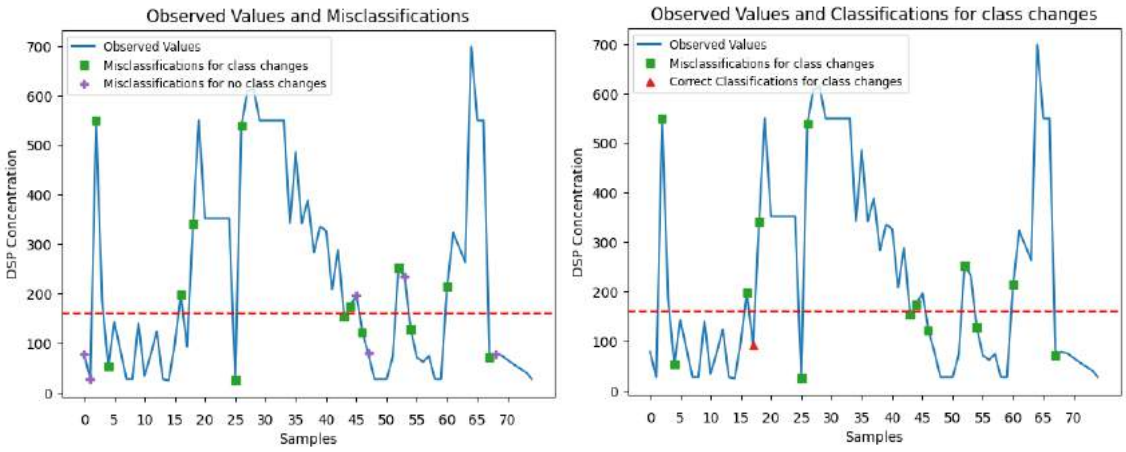


(a) Misclassifications.



(b) Class Changing samples.

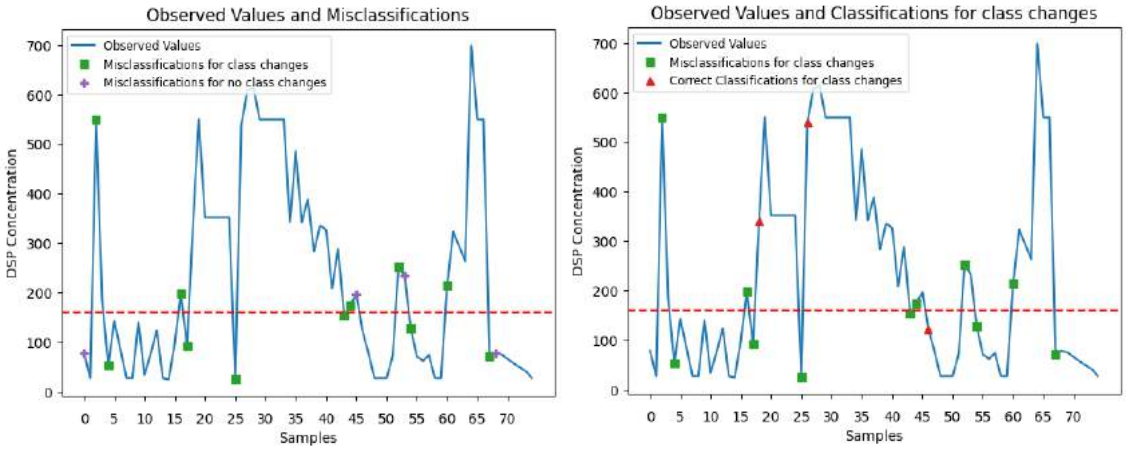
Figure C.12: Misclassifications and Class changing samples in RIAV1 for the RIAV1-D model.



(a) Misclassifications.

(b) Class Changing samples.

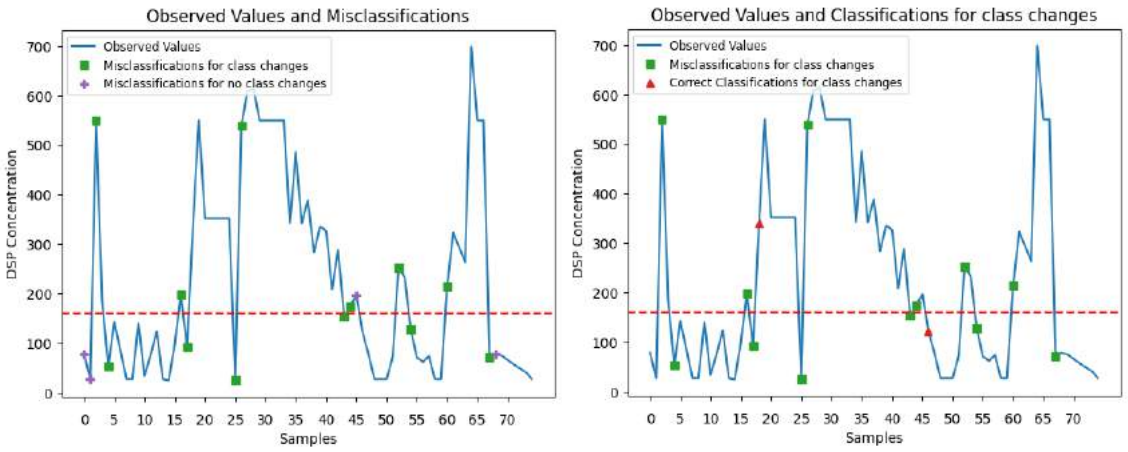
Figure C.13: Misclassifications and Class changing samples in RIAV1 for the RIAV1-DM model.



(a) Misclassifications.

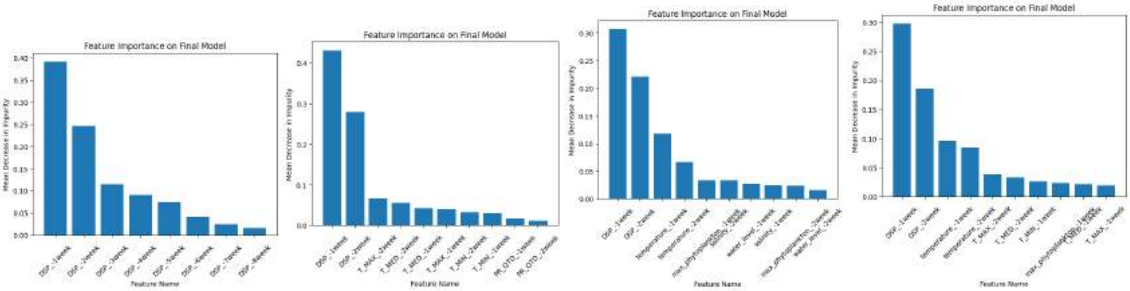
(b) Class Changing samples.

Figure C.14: Misclassifications and Class changing samples in RIAV1 for the RIAV1-DH model.



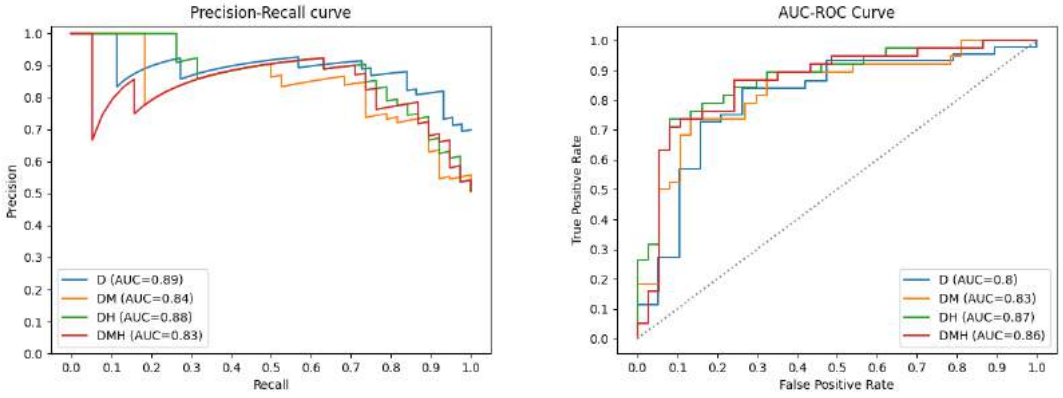
(a) Misclassifications. (b) Class Changing samples.

Figure C.15: Misclassifications and Class changing samples in RIAV1 for the RIAV1-DMH model.



(a) RIAV1-D. (b) RIAV1-DM. (c) RIAV1-DH. (d) RIAV1-DHM.

Figure C.16: RIAV1 RF Classification feature importance.



(a) PR-REC. (b) ROC-AUC.

Figure C.17: RIAV1 Triângulo RF Classification models curves.

Table C.3: RF Classification metrics for RIAV1 Triângulo

Set	Metric	Dataset			
		DSP	Meteo	Hydrodynamic	Meteo and Hydro
Train	Balanced Accuracy	0.8909	**0.9291**	0.9016	0.8889
	Average Precision	0.9641	**0.9868**	0.9653	0.9636
	F1 Macro	0.8931	**0.9296**	0.9022	0.8880
	Recall	0.8696	**0.9321**	0.9247	0.9220
Validation	Balanced Accuracy	**0.8775**	0.8472	0.8478	0.8458
	Average Precision	0.8969	**0.9089**	0.9018	0.8939
	F1 Macro	**0.8650**	0.8434	0.8317	0.8271
	Recall	**0.8765**	0.8237	0.8181	0.8299
Test	Balanced Accuracy	0.7362	0.7464	**0.7998**	0.7863
	Average Precision	**0.8899**	0.8360	0.8800	0.8336
	F1 Macro	0.7362	0.7465	**0.7999**	0.7863
	Recall	**0.8409**	0.7632	0.8158	0.8158

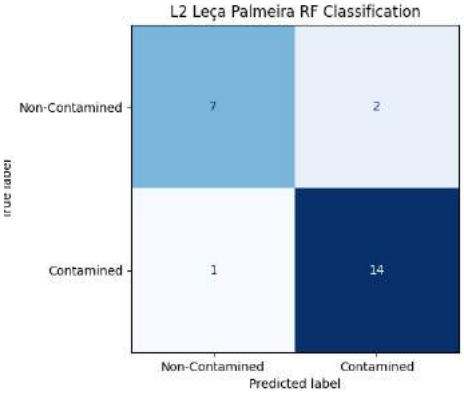
C.1.1.4 RF Classification Upwelling L2 Leça da Palmeira

Table C.4: RF Upwelling Classification metrics for L2 Leça da Palmeira

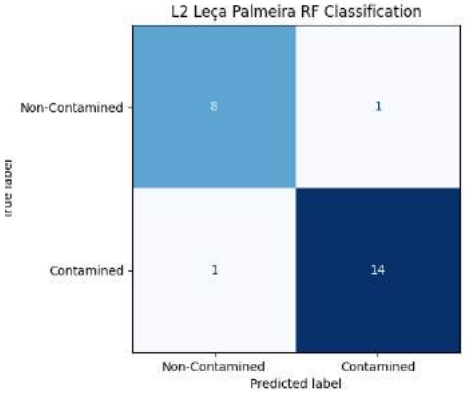
Set	Metric	Dataset			
		D	DU	DM	DMU
Train	Balanced Accuracy	**0.9076**	0.6988	0.9017	0.8686
	Average Precision	0.9947	0.9945	**0.9989**	0.9975
	F1 Macro	0.9065	0.6838	**0.9177**	0.9005
	Recall	0.9629	0.9907	0.9760	**0.9946**
Validation	Balanced Accuracy	**0.8797**	0.7126	0.7065	0.6688
	Average Precision	0.9609	0.9166	**0.9652**	0.8922
	F1 Macro	**0.8719**	0.6773	0.6765	0.6373
	Recall	0.9023	0.9722	**0.9792**	0.9722
Test	Balanced Accuracy	**0.9643**	0.8556	0.8556	0.9111
	Average Precision	0.9786	**0.9844**	0.9750	0.9810
	F1 Macro	**0.9521**	0.8634	0.8634	0.9111
	Recall	0.9286	**0.9333**	**0.9333**	**0.9333**

(a) L2-D.

(b) L2-DU.

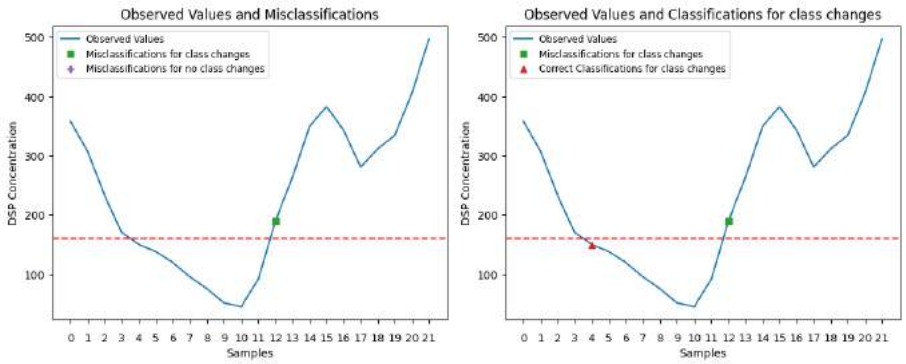


(c) L2-DM.



(d) L2-DMU.

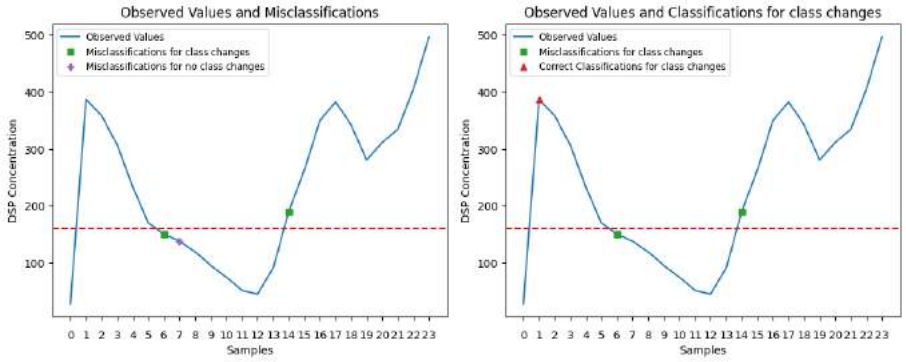
Figure C.18: RF Classification Confusion Matrices for L2 Leça da Palmeira.



(a) Missing distribution.

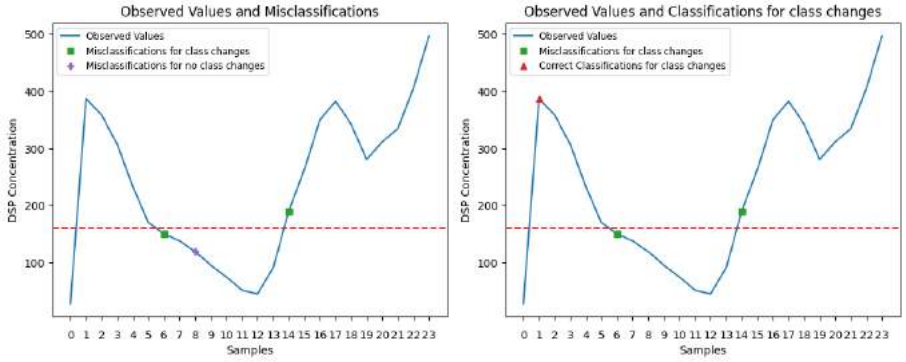
(b) Missing periods.

Figure C.19: L2-UP-D class changes and misclassifications.



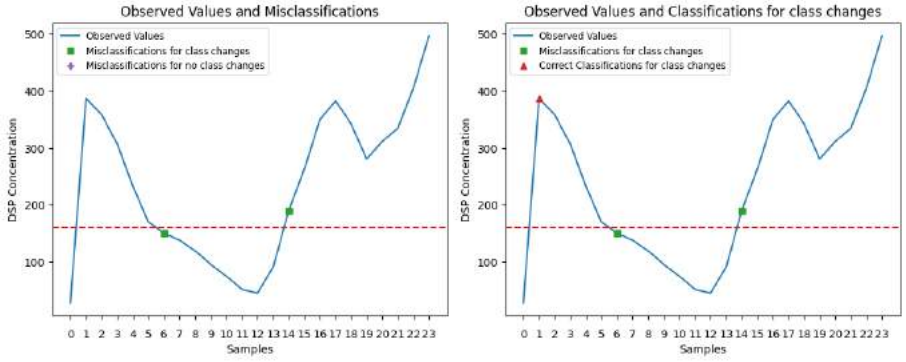
(a) Missing distribution. (b) Missing periods.

Figure C.20: L2-UP-DU class changes and misclassifications.



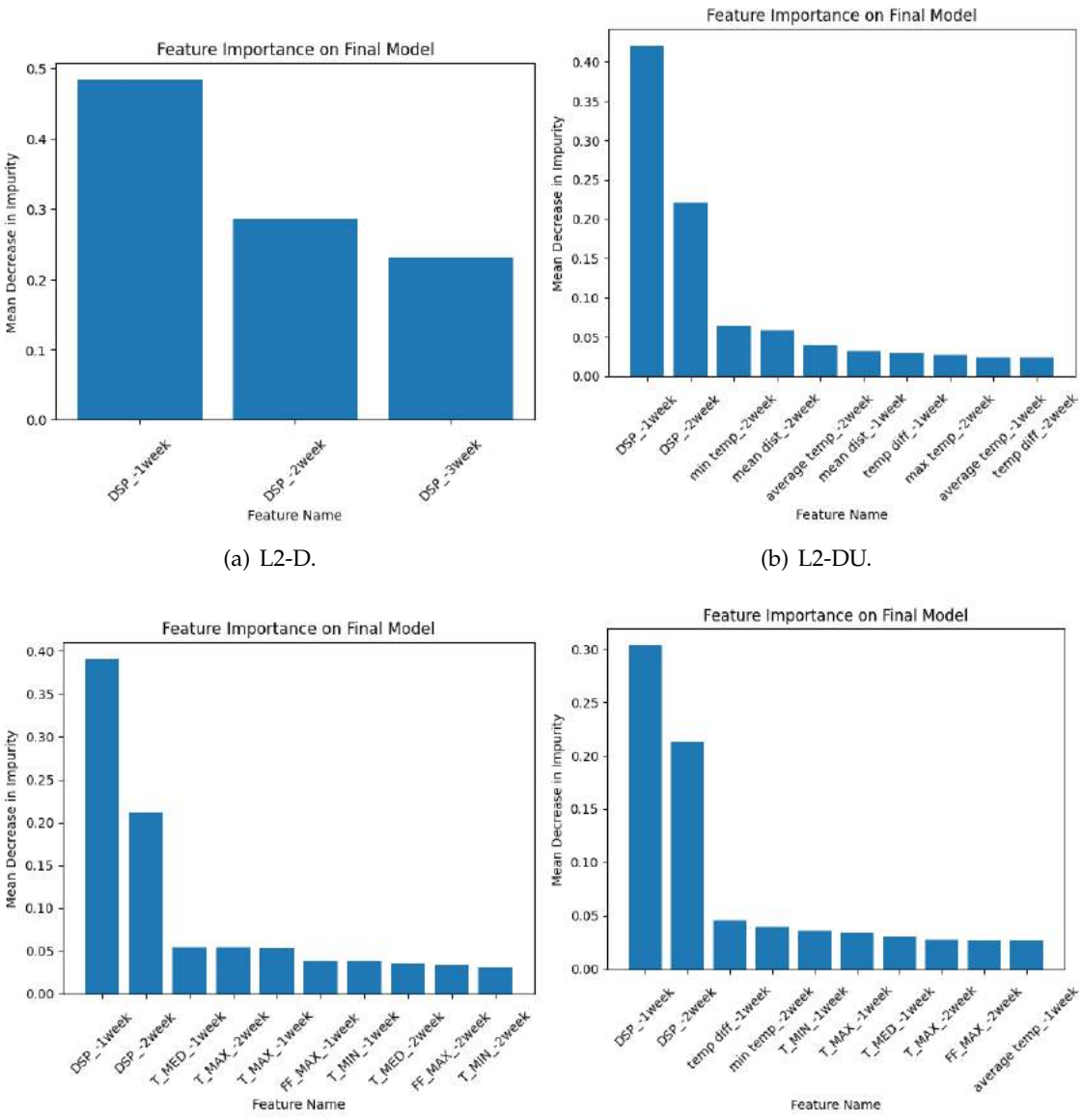
(a) Missing distribution. (b) Missing periods.

Figure C.21: L2-UP-DM class changes and misclassifications.



(a) Missing distribution. (b) Missing periods.

Figure C.22: L2-UP-DMU class changes and misclassifications.



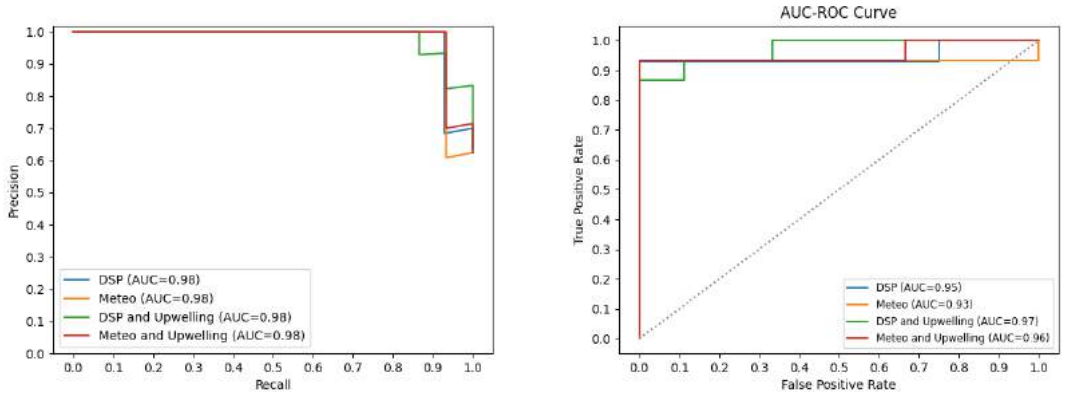
(a) L2-D.

(b) L2-DU.

(c) L2-DM.

(d) L2-DMU.

Figure C.23: L2 Leça Palmeira RF Classification feature importance.



(a) PR-REC. (b) AUC-ROC.

Figure C.24: L2 Leça Palmeira RF Classification models curves.

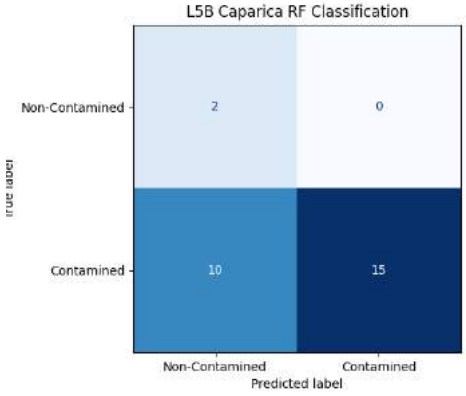
C.1.1.5 RF Classification Upwelling L5b Caparica

Table C.5: RF Upwelling Classification metrics for L5b Caparica

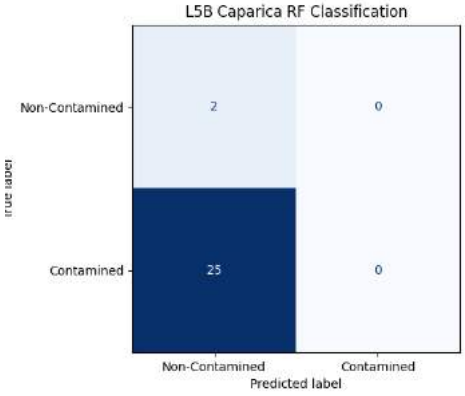
Set	Metric	Dataset			
		D	DU	DM	DMU
Train	Balanced Accuracy	0.9841	0.8238	0.8110	**0.9905**
	Average Precision	0.9987	0.9992	0.9992	**0.9997**
	F1 Macro	0.9856	0.7784	0.7685	**0.9905**
	Recall	**0.9938**	**0.9938**	**0.9938**	**0.9938**
Validation	Balanced Accuracy	0.9173	**0.9438**	0.9292	0.8962
	Average Precision	0.9102	0.9306	0.9245	**0.9447**
	F1 Macro	0.9081	**0.9332**	0.9212	0.8807
	Recall	**0.9722**	**0.9722**	**0.9722**	**0.9722**
Test	Balanced Accuracy	0.6900	0.4200	**0.8000**	0.5000
	Average Precision	**0.9790**	0.9496	0.9757	0.9682
	F1 Macro	**0.6250**	0.4375	0.5179	0.0690
	Recall	**0.8800**	0.8400	0.6000	0.000

(a) L5b-D.

(b) L5b-DU.

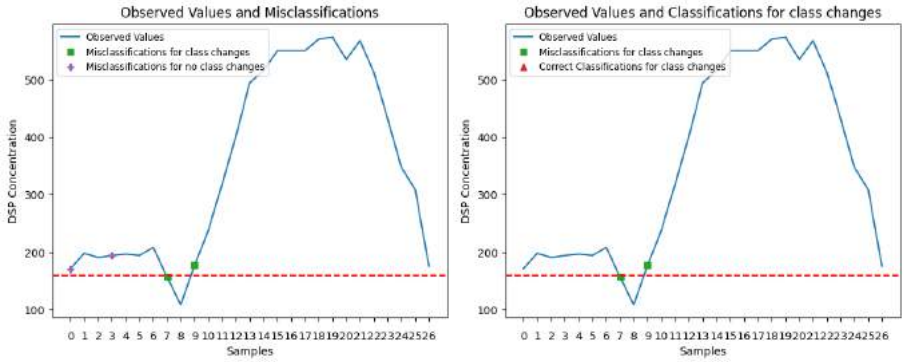


(c) L5b-DM.



(d) L5b-DMU.

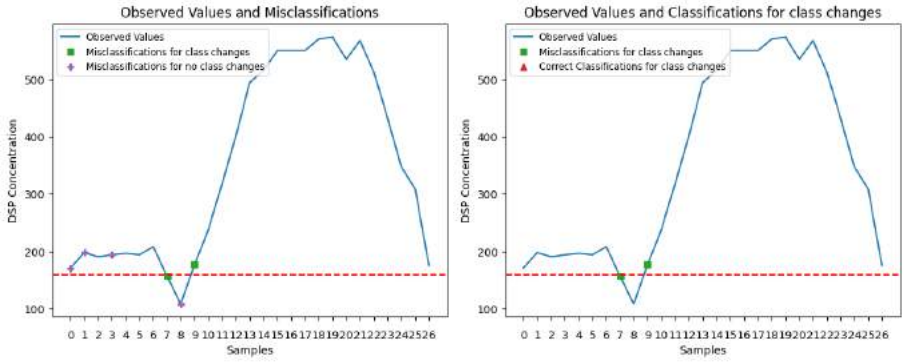
Figure C.25: RF Classification Confusion Matrices for L5b Caparica.



(a) Missing distribution.

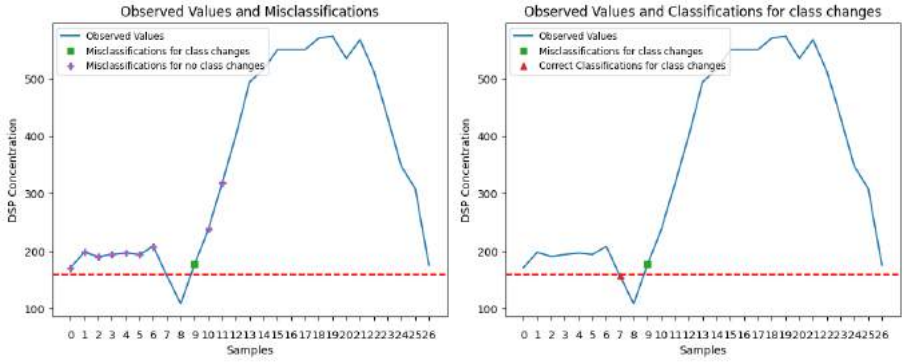
(b) Missing periods.

Figure C.26: L5b-UP-D class changes and misclassifications.



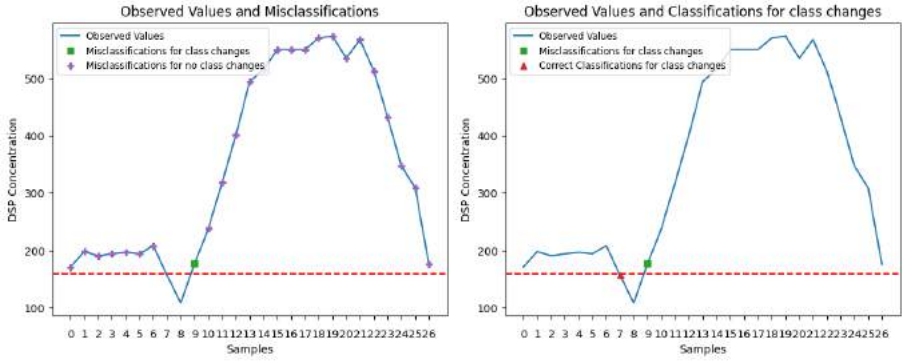
(a) Missing distribution. (b) Missing periods.

Figure C.27: L5b-UP-DU class changes and misclassifications.



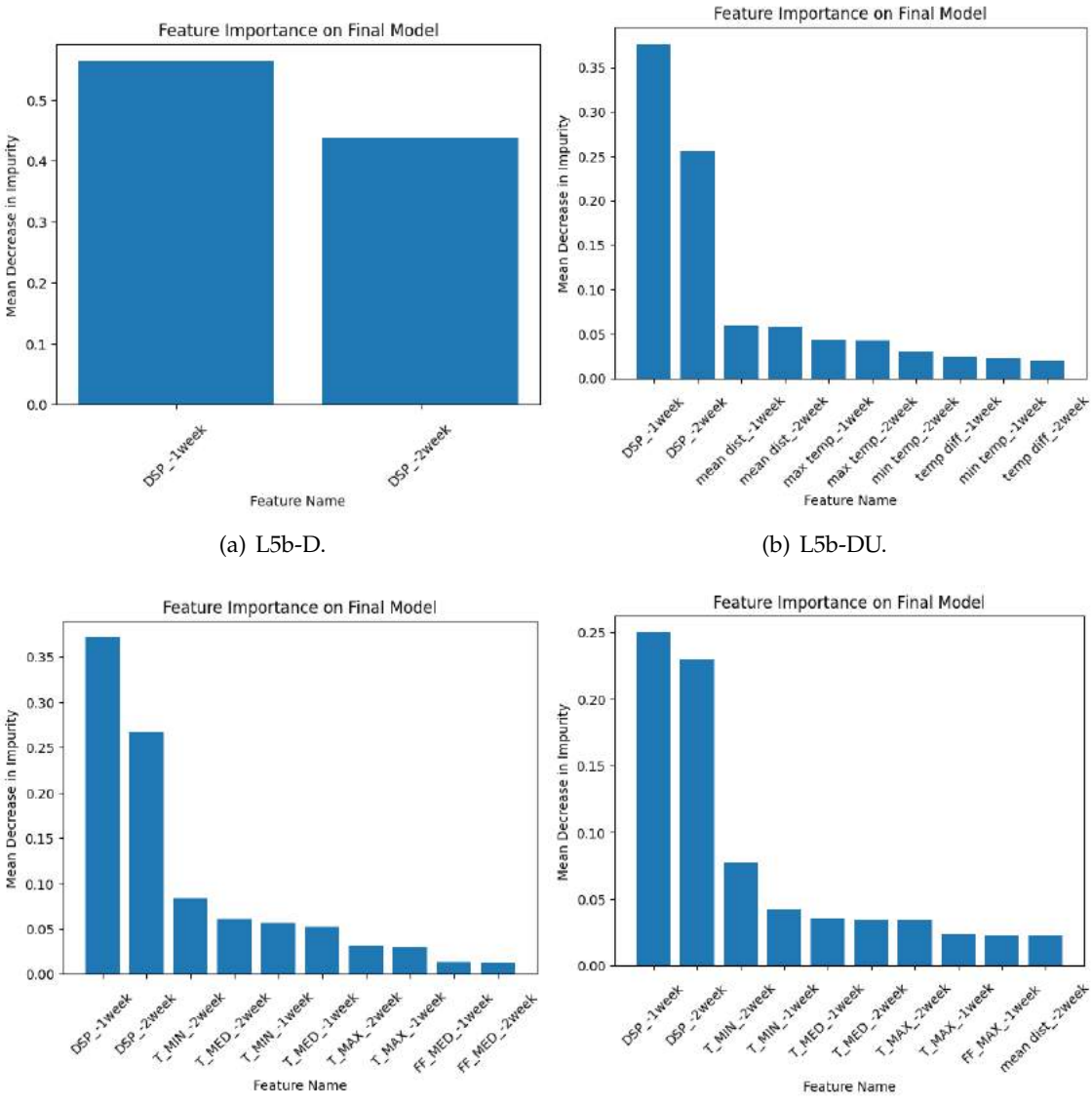
(a) Missing distribution. (b) Missing periods.

Figure C.28: L5b-UP-DM class changes and misclassifications.



(a) Missing distribution. (b) Missing periods.

Figure C.29: L5b-UP-DMU class changes and misclassifications.



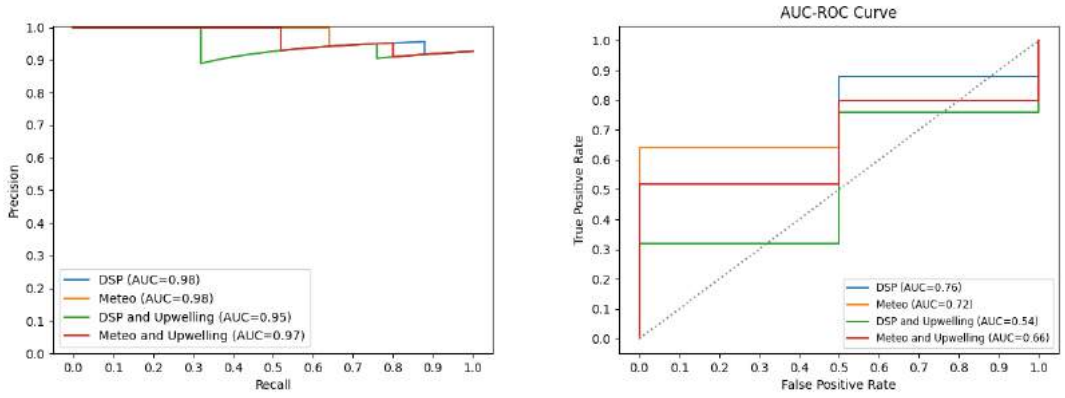
(a) L5b-D.

(b) L5b-DU.

(c) L5b-DM.

(d) L5b-DMU.

Figure C.30: L5b Caparica RF Classification feature importance.



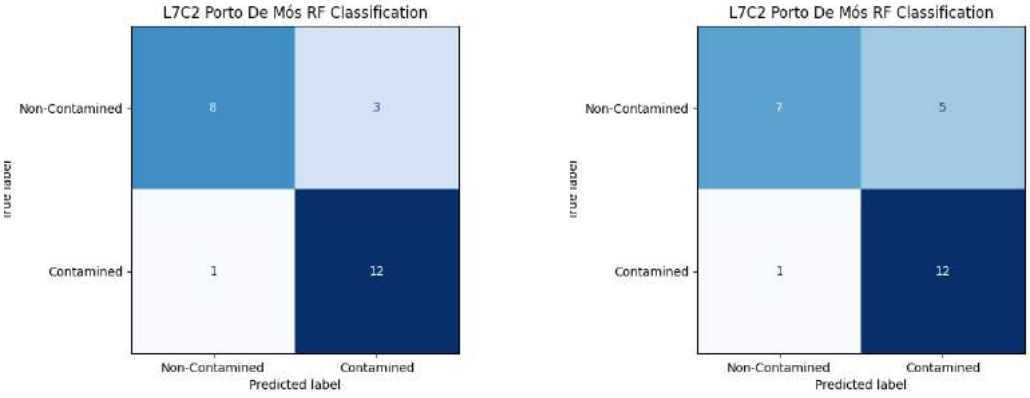
(a) PR-REC. (b) AUC-ROC.

Figure C.31: L5b Caparica RF Classification models curves.

C.1.1.6 RF Classification Upwelling L7c2 Porto de Mós

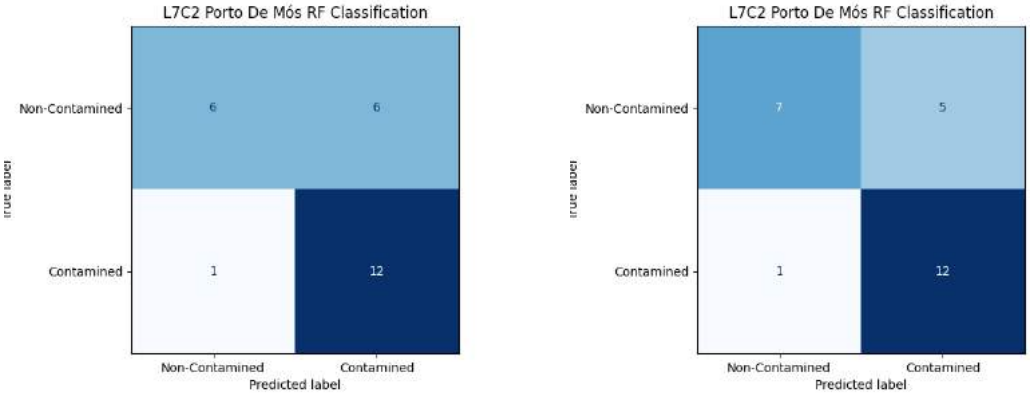
Table C.6: RF Upwelling Classification metrics for L7c2 Porto de Mós

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	Balanced Accuracy	0.8991	0.8909	0.8865	**0.9334**
	Average Precision	0.9624	0.9288	0.9667	**0.9919**
	F1 Macro	0.8983	0.8950	0.8890	**0.9344**
	Recall	0.8657	0.8274	0.8256	**0.9124**
Validation	Balanced Accuracy	**0.9296**	0.8392	0.8806	0.7611
	Average Precision	**0.9930**	0.8634	0.9001	0.7278
	F1 Macro	**0.9460**	0.8445	0.8948	0.7755
	Recall	**0.9333**	0.7489	0.8458	0.6068
Test	Balanced Accuracy	**0.8252**	0.7532	0.7115	0.7532
	Average Precision	0.9626	0.9607	0.9648	**0.9731**
	F1 Macro	**0.8286**	0.7500	0.7029	0.7500
	Recall	**0.9231**	**0.9231**	**0.9231**	**0.9231**



(a) L7c2-D.

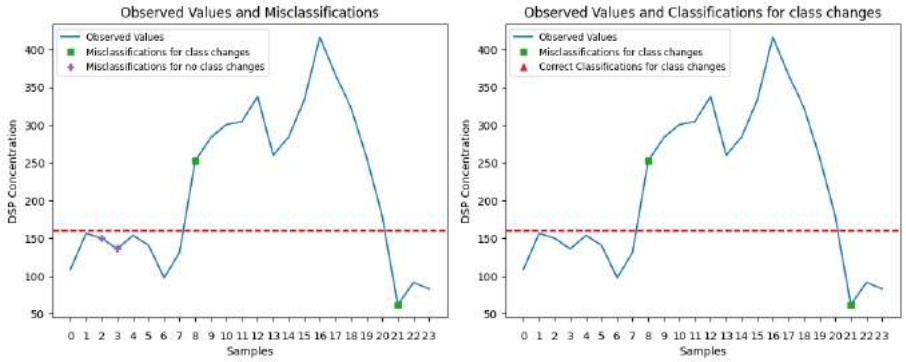
(b) L7c2-DU.



(c) L7c2-DM.

(d) L7c2-DMU.

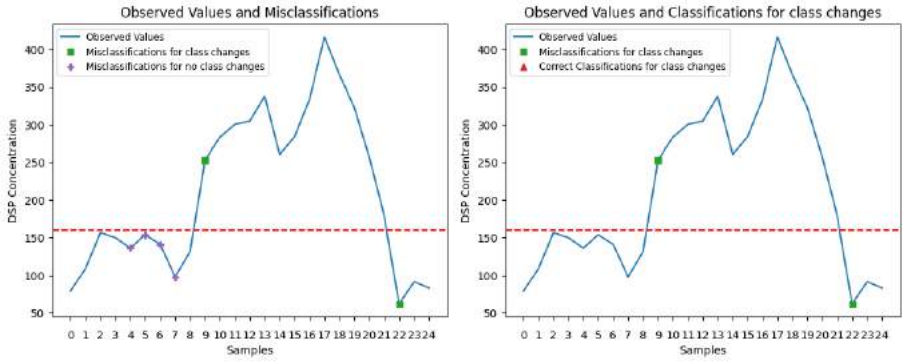
Figure C.32: RF Classification Confusion Matrices for L7c2 Porto de Mós.



(a) Missing distribution.

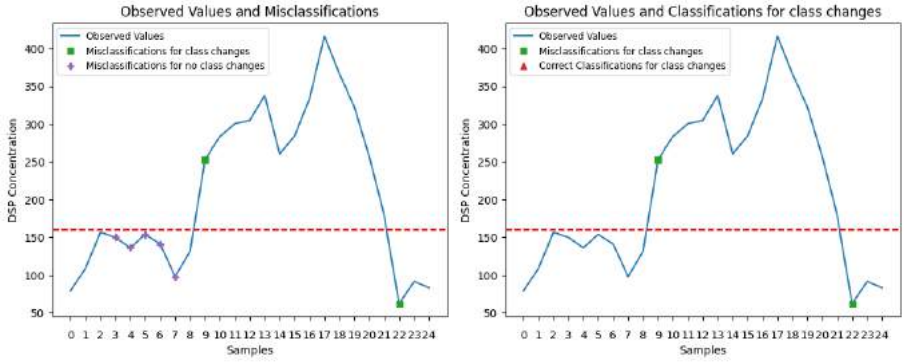
(b) Missing periods.

Figure C.33: L7c2-UP-D class changes and misclassifications.



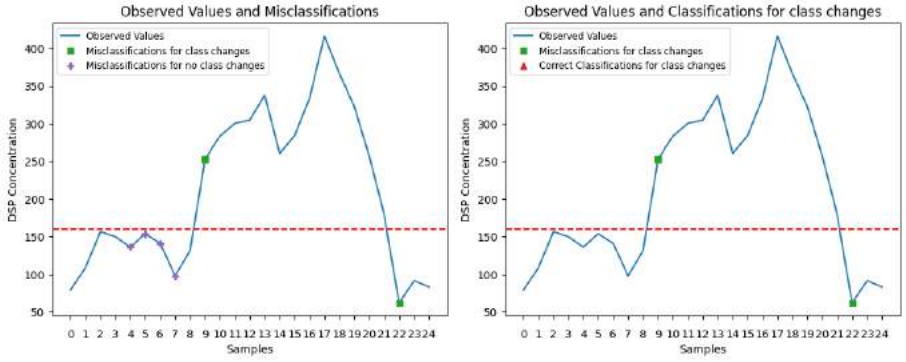
(a) Missing distribution. (b) Missing periods.

Figure C.34: L7c2-UP-DU class changes and misclassifications.



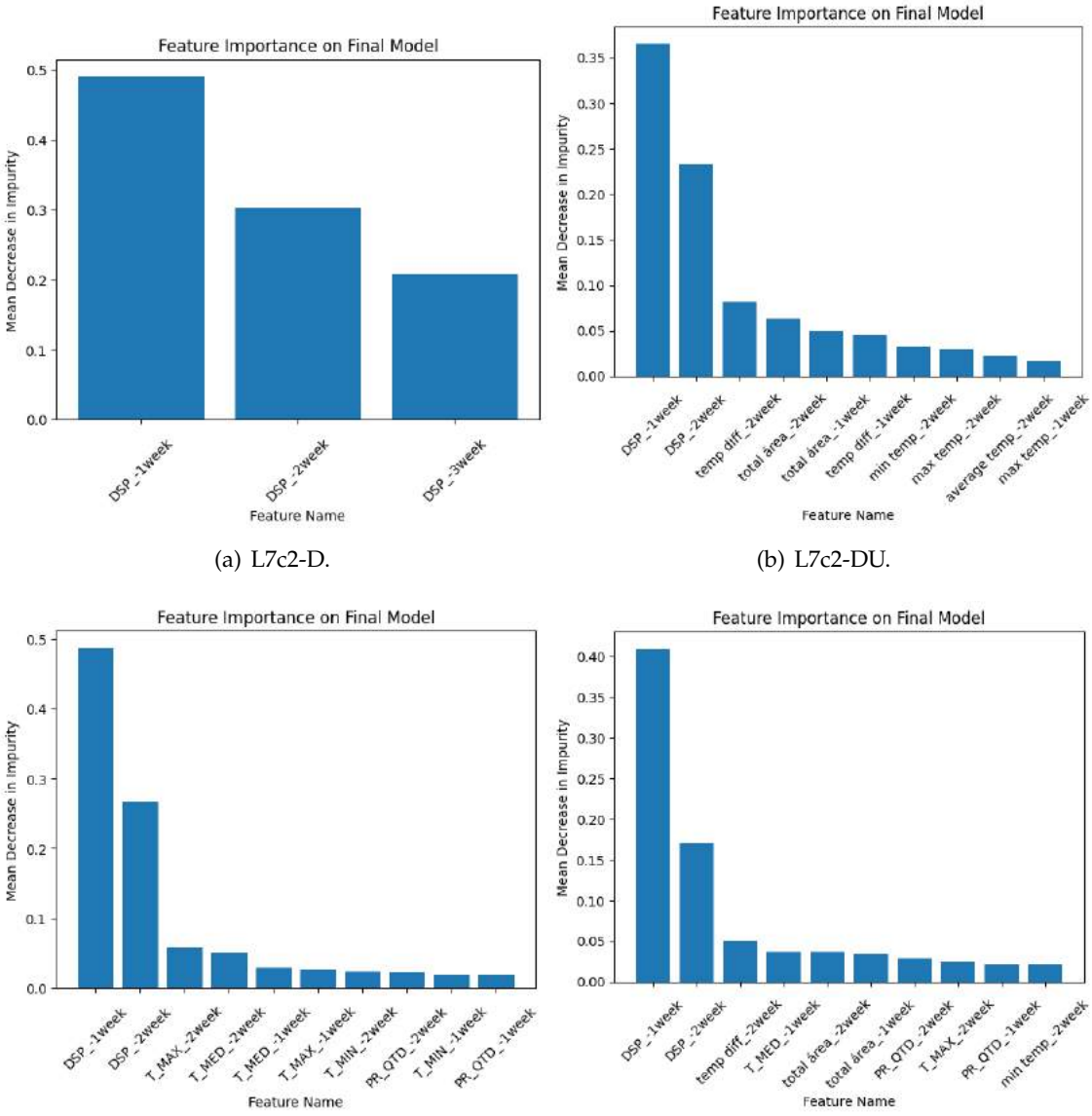
(a) Missing distribution. (b) Missing periods.

Figure C.35: L7c2-UP-DM class changes and misclassifications.



(a) Missing distribution. (b) Missing periods.

Figure C.36: L7c2-UP-DMU class changes and misclassifications.



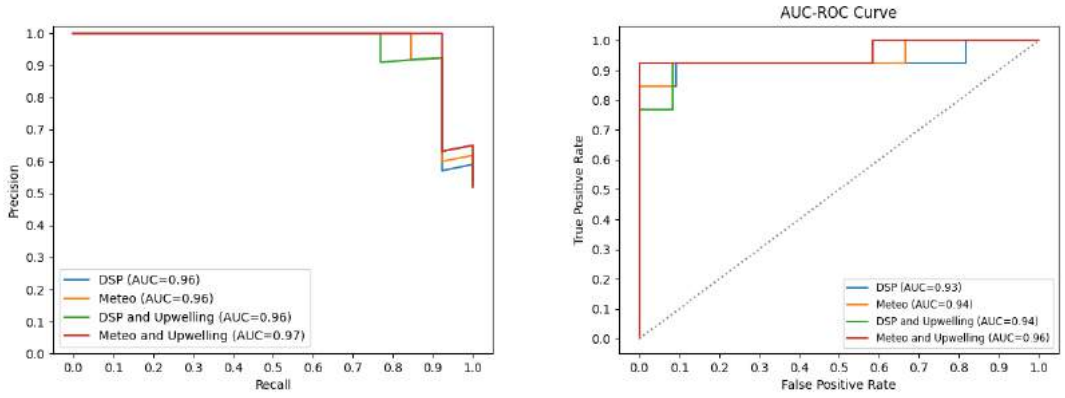
(a) L7c2-D.

(b) L7c2-DU.

(c) L7c2-DM.

(d) L7c2-DMU.

Figure C.37: L7c2 Porto de Mós RF Classification feature importance.



(a) PR-REC. (b) AUC-ROC.

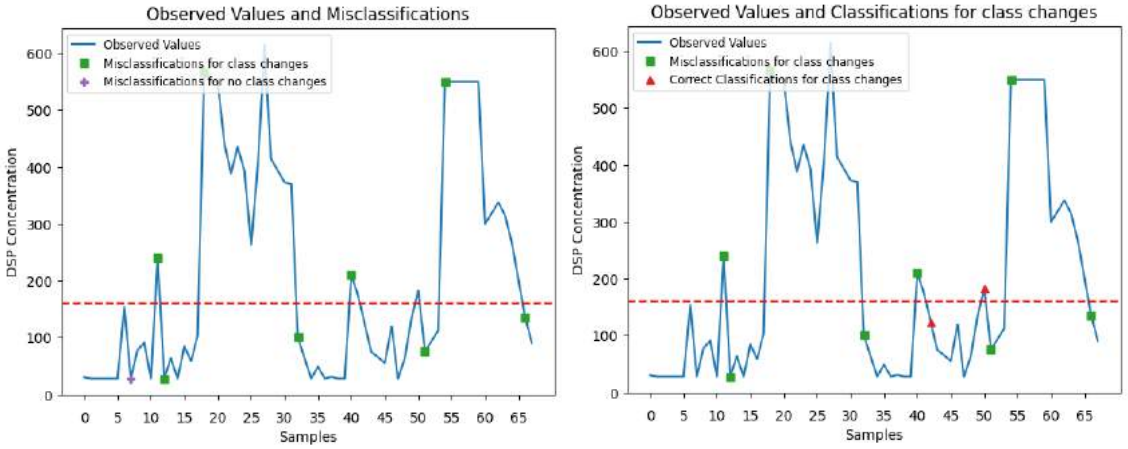
Figure C.38: L7c2 Porto de Mós RF Classification models curves.

C.1.2 Support Vector Machine

C.1.2.1 SVM Classification L2 Leça da Palmeira

(a) L2-D. (b) L2-DM.

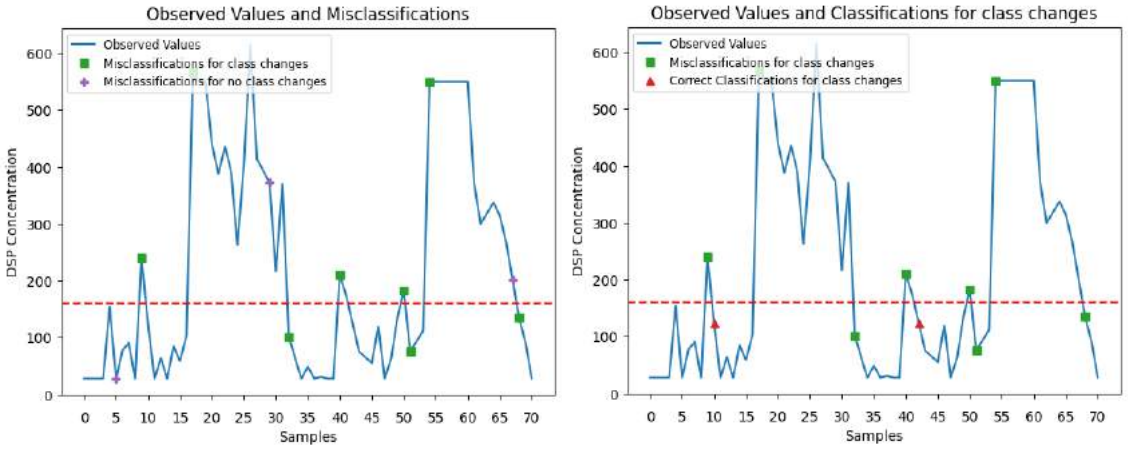
Figure C.39: SVM Classification Confusion Matrices for L2 Leça da Palmeira.



(a) Misclassifications.

(b) Class Changing samples.

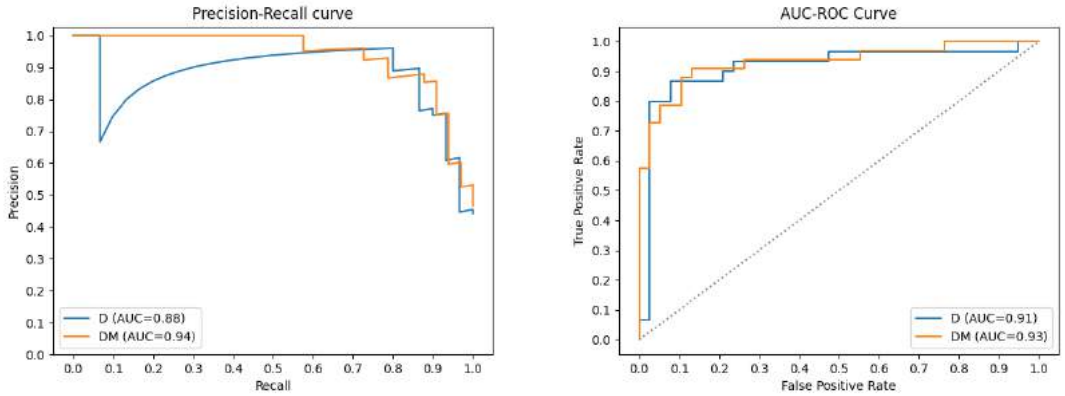
Figure C.40: Misclassifications and Class changing samples for the L2-D model.



(a) Misclassifications.

(b) Class Changing samples.

Figure C.41: Misclassifications and Class changing samples for the L2-DM model.



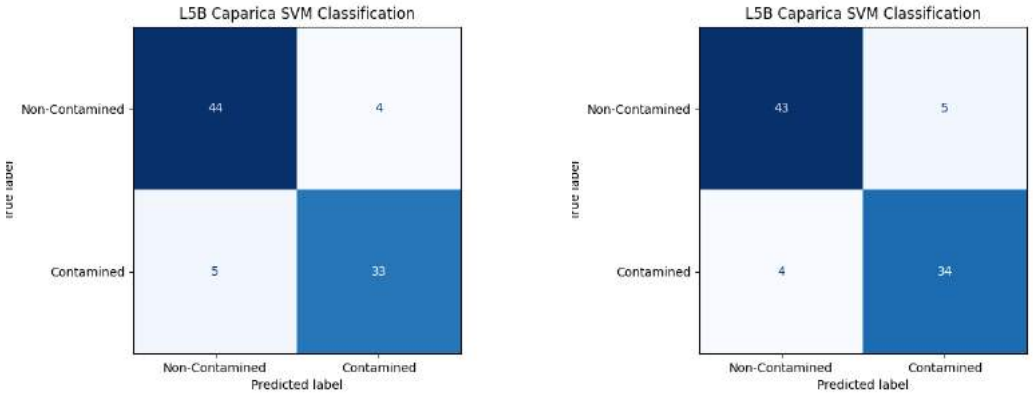
(a) PR-REC. (b) AUC-ROC.

Figure C.42: L2 Leça da Palmeira SVM Classification models curves.

Table C.7: SVM Classification metrics for L2 Leça da Palmeira

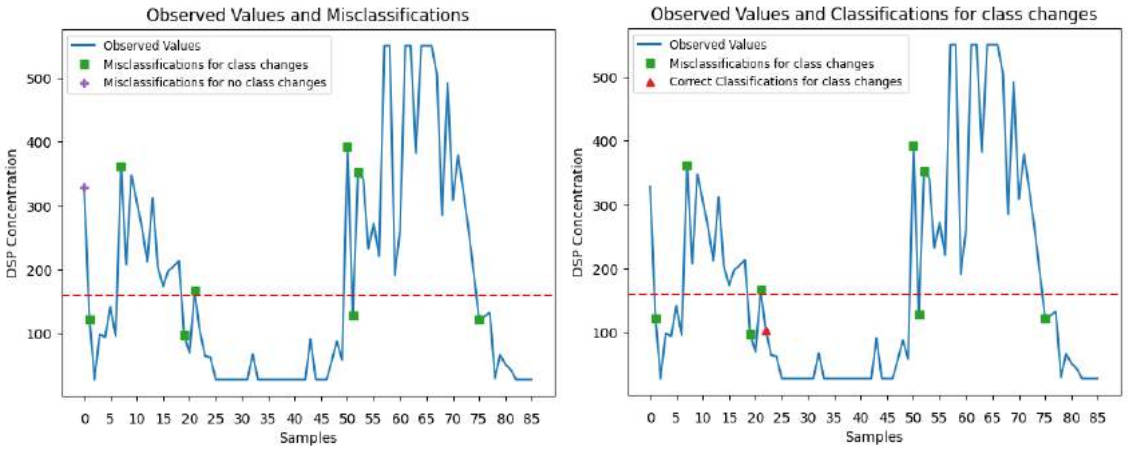
Set	Metric	Dataset	
		D	DM
Train	Balanced Accuracy	0.8710	**0.9115**
	Average Precision	0.9162	**0.9501**
	F1 Macro	0.8731	**0.9121**
	Recall	0.8190	**0.8777**
Validation	Balanced Accuracy	**0.7969**	0.7758
	Average Precision	**0.8407**	0.7822
	F1 Macro	**0.8043**	0.7730
	Recall	**0.7157**	0.6881
Test	Balanced Accuracy	**0.8675**	0.8413
	Average Precision	0.8803	**0.9397**
	F1 Macro	**0.8662**	0.8431
	Recall	**0.8667**	0.7879

C.1.2.2 SVM Classification L5b Caparica



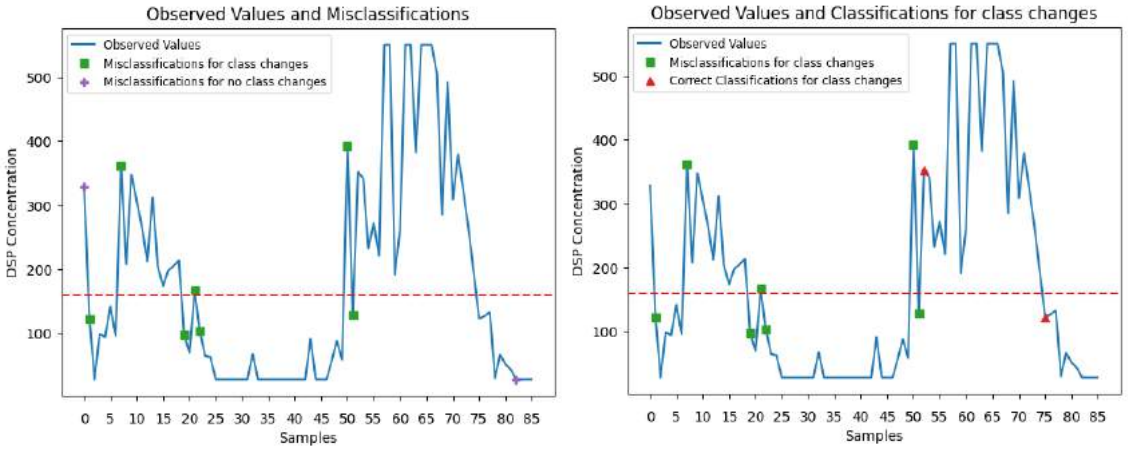
(a) L5b-D. (b) L5b-DM.

Figure C.43: SVM Classification Confusion Matrices for L5b Caparica.



(a) Misclassifications. (b) Class Changing samples.

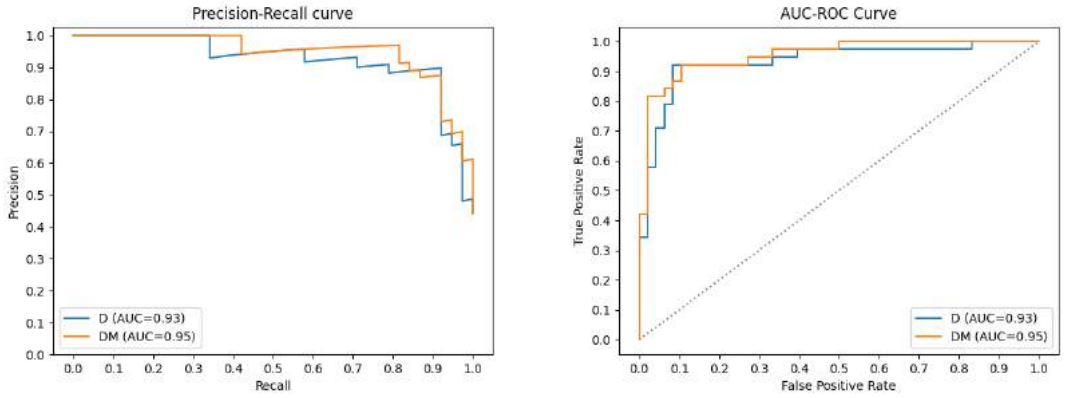
Figure C.44: Misclassifications and Class changing samples for the L5b-D SVM model.



(a) Misclassifications.

(b) Class Changing samples.

Figure C.45: Misclassifications and Class changing samples for the L5b-DM SVM model.



(a) PR-REC.

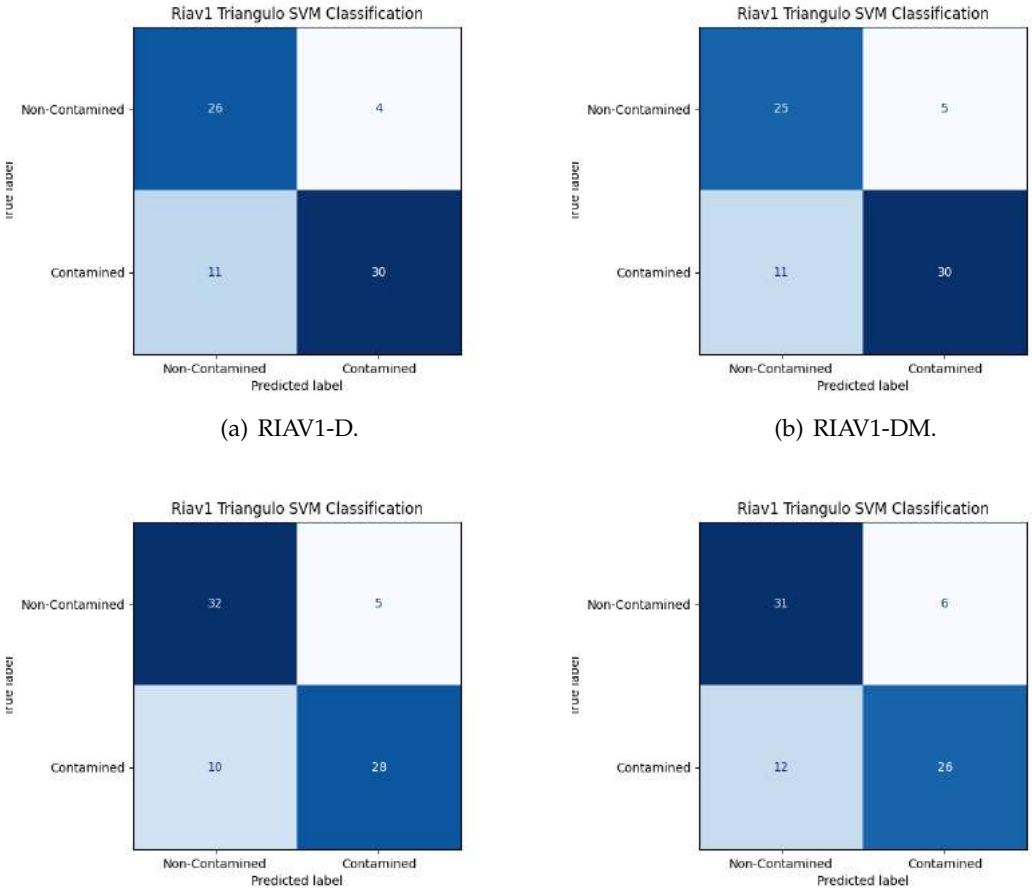
(b) ROC-AUC.

Figure C.46: L5b Caparica SVM Classification models curves.

Table C.8: SVM Classification metrics for L5b Caparica

Set	Metric	Dataset	
		D	DM
Train	Balanced Accuracy	**0.9105**	0.9037
	Average Precision	**0.9372**	0.9265
	F1 Macro	**0.9133**	0.9090
	Recall	**0.8768**	0.8561
Validation	Balanced Accuracy	**0.8202**	0.8019
	Average Precision	0.7955	**0.8859**
	F1 Macro	**0.8233**	0.8083
	Recall	**0.7630**	0.6981
Test	Balanced Accuracy	0.8925	**0.8953**
	Average Precision	0.9258	**0.9468**
	F1 Macro	0.8936	**0.8942**
	Recall	0.8684	**0.8947**

C.1.2.3 SVM Classification RIAV1 Triângulo



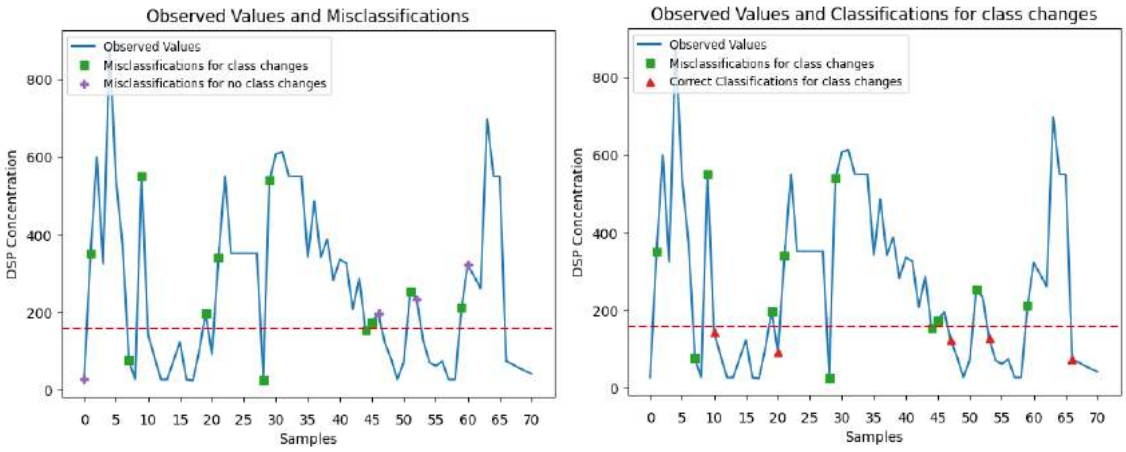
(a) RIAV1-D.

(b) RIAV1-DM.

(c) RIAV1-DH.

(d) RIAV1-DHM.

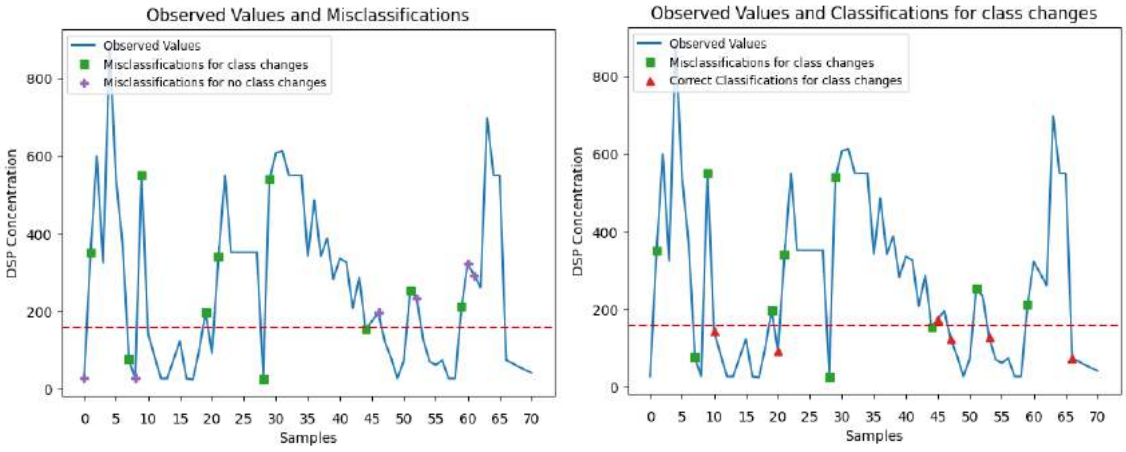
Figure C.47: SVM Classification Confusion Matrices for RIAV1 Triângulo.



(a) Misclassifications.

(b) Class Changing samples.

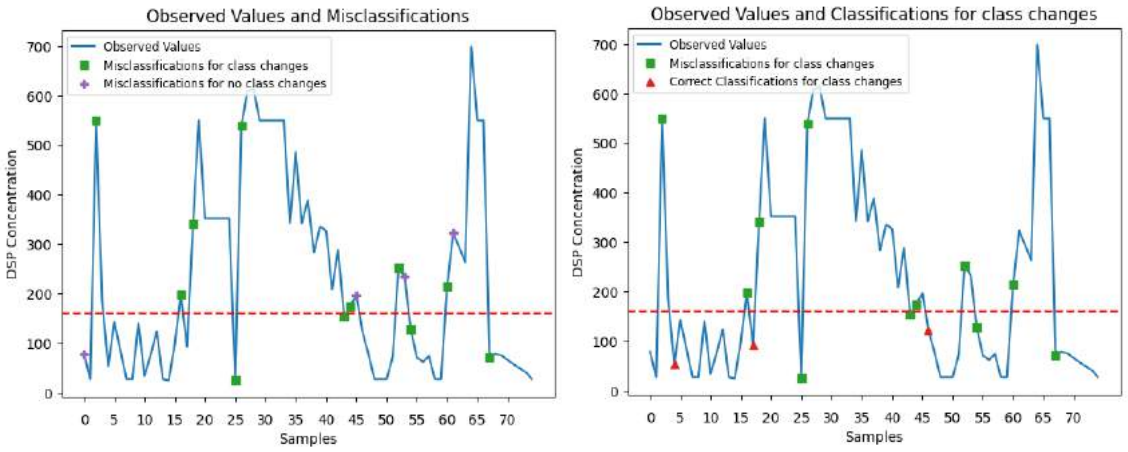
Figure C.48: Misclassifications and Class changing samples in RIAV1 for the RIAV1-D model.



(a) Misclassifications.

(b) Class Changing samples.

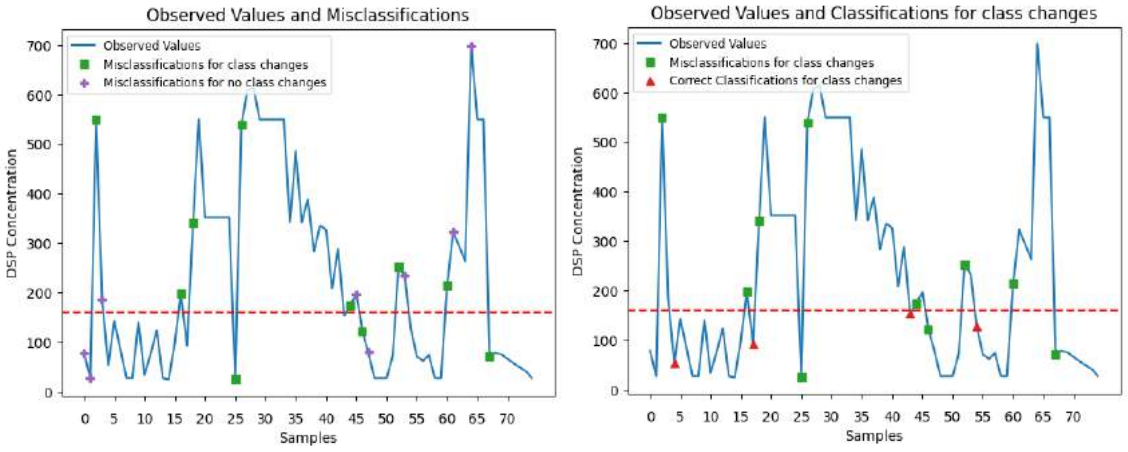
Figure C.49: Misclassifications and Class changing samples in RIAV1 for the RIAV1-DM model.



(a) Misclassifications.

(b) Class Changing samples.

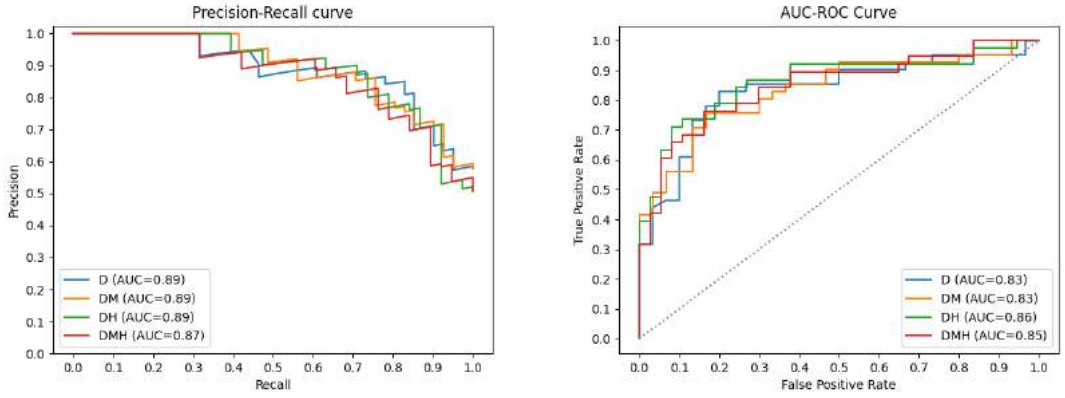
Figure C.50: Misclassifications and Class changing samples in RIAV1 for the RIAV1-DH model.



(a) Misclassifications.

(b) Class Changing samples.

Figure C.51: Misclassifications and Class changing samples in RIAV1 for the RIAV1-DMH model.



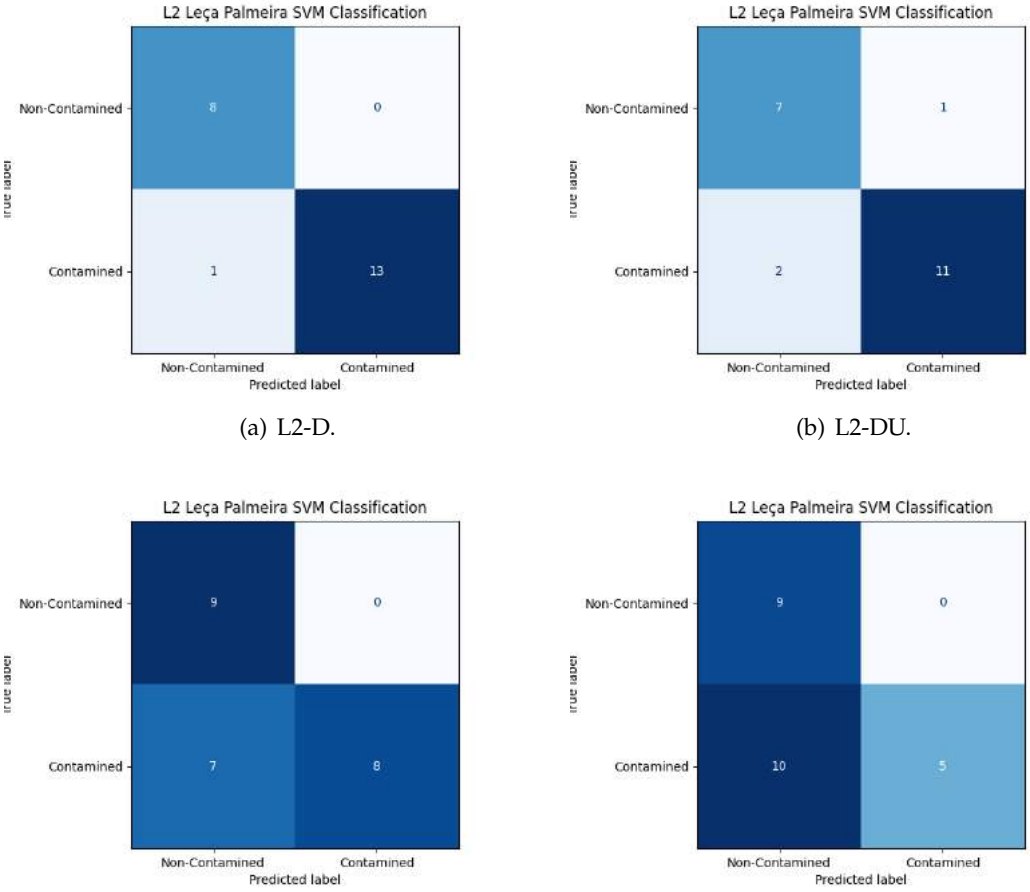
(a) PR-REC. (b) ROC-AUC.

Figure C.52: RIAV1 Triângulo SVM Classification models curves.

Table C.9: SVM Classification metrics for RIAV1 Triângulo

Set	Metric	Dataset			
		D	DM	DH	DMH
Train	Balanced Accuracy	**0.9082**	0.8353	0.8931	0.8850
	Average Precision	0.9546	0.9199	0.9508	**0.9621**
	F1 Macro	**0.9129**	0.8370	0.8915	0.8841
	Recall	0.8308	0.7575	**0.8909**	0.8884
Validation	Balanced Accuracy	**0.8709**	0.8680	0.8183	0.8240
	Average Precision	**0.9325**	0.9211	0.8973	0.8665
	F1 Macro	**0.8728**	0.8591	0.8132	0.8030
	Recall	0.8022	**0.8117**	0.7709	0.7736
Test	Balanced Accuracy	0.7992	0.7825	**0.8009**	0.7610
	Average Precision	0.8875	**0.8925**	0.8907	0.8726
	F1 Macro	0.7881	0.7735	**0.7994**	0.7589
	Recall	0.7317	0.7317	0.7368	0.6842

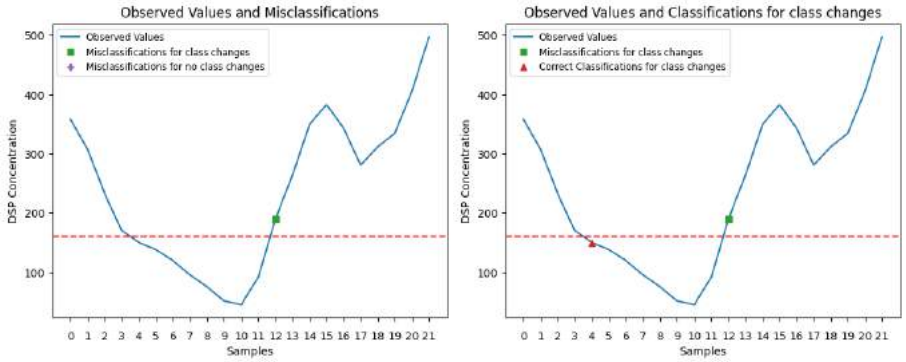
C.1.2.4 SVM Classification Upwelling L2 Leça da Palmeira



(a) L2-D. (b) L2-DU.

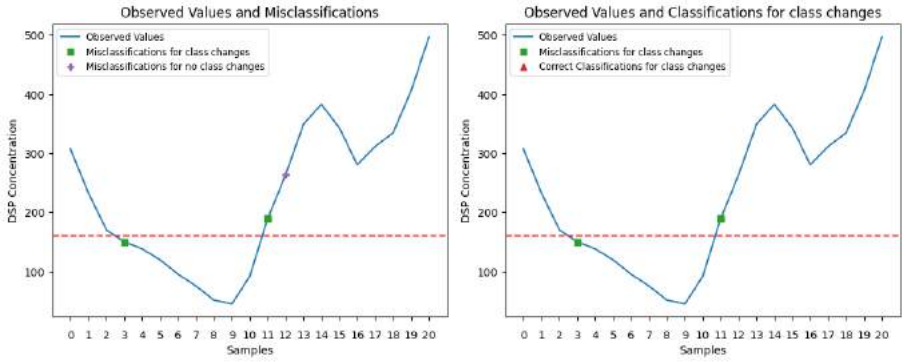
(c) L2-DM. (d) L2-DMU.

Figure C.53: SVM Classification Confusion Matrices for L2 Leça da Palmeira.



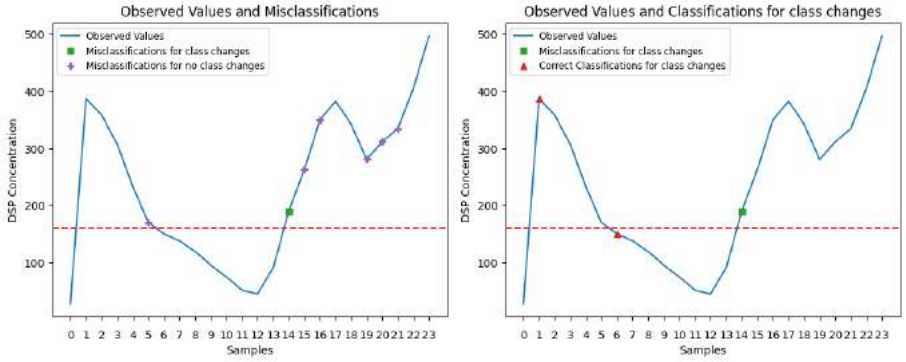
(a) Missing distribution. (b) Missing periods.

Figure C.54: SVM L2-UP-D class changes and misclassifications.



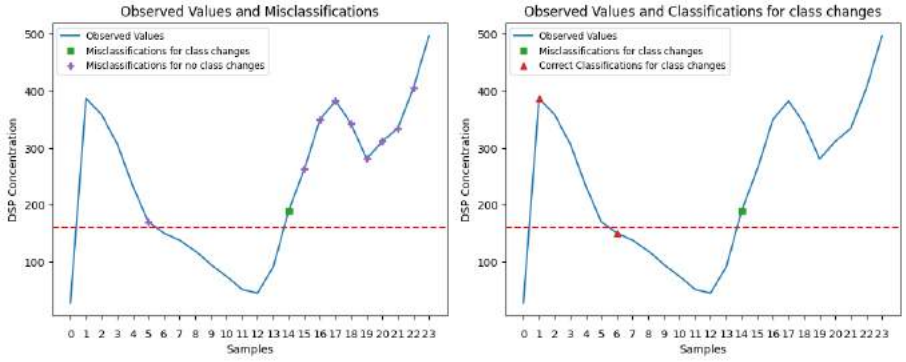
(a) Missing distribution. (b) Missing periods.

Figure C.55: SVM L2-UP-DU class changes and misclassifications.



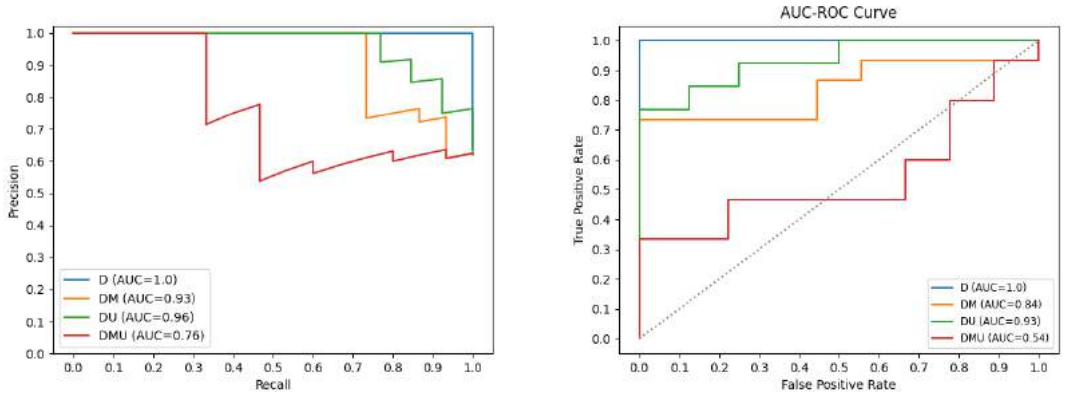
(a) Missing distribution. (b) Missing periods.

Figure C.56: SVM L2-UP-DM class changes and misclassifications.



(a) Missing distribution. (b) Missing periods.

Figure C.57: SVM L2-UP-DMU class changes and misclassifications.



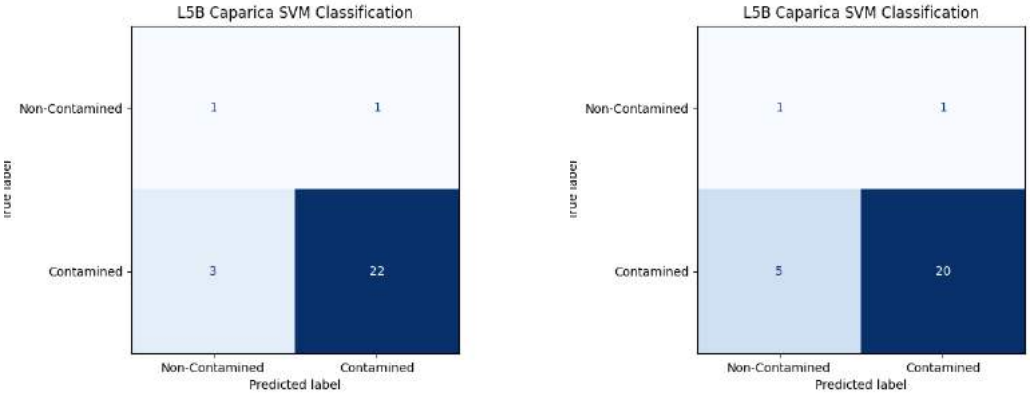
(a) PR-REC. (b) AUC-ROC.

Figure C.58: L2 Leça Palmeira SVM Classification models curves.

Table C.10: SVM Upwelling Classification metrics for L2 Leça da Palmeira

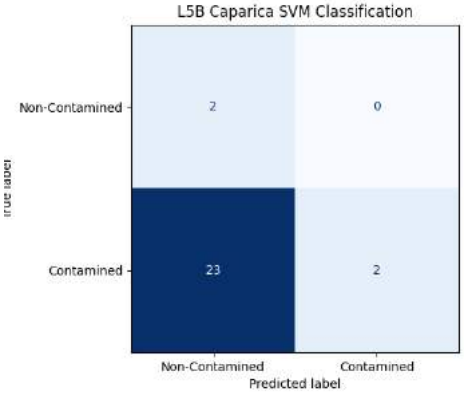
Set	Metric	Dataset			
		D	DU	DM	DMU
Train	Balanced Accuracy	**0.8908**	0.8877	0.8513	0.8674
	Average Precision	0.9930	0.9941	0.9925	**0.9951**
	F1 Macro	0.7898	**0.8012**	0.7358	0.7577
	Recall	**0.7816**	0.7755	0.7025	0.7347
Validation	Balanced Accuracy	**0.9422**	0.8812	0.8697	0.8531
	Average Precision	**0.9878**	0.9591	0.9620	0.9397
	F1 Macro	**0.9178**	0.8403	0.8533	0.8361
	Recall	0.8844	0.8636	0.8889	**0.9306**
Test	Balanced Accuracy	**0.9643**	0.8606	0.7667	0.6667
	Average Precision	**1.0000**	0.9645	0.9251	0.7607
	F1 Macro	**0.9521**	0.8518	0.7078	0.5714
	Recall	**0.9286**	0.8462	0.5333	0.3333

C.1.2.5 SVM Classification Upwelling L5b Caparica

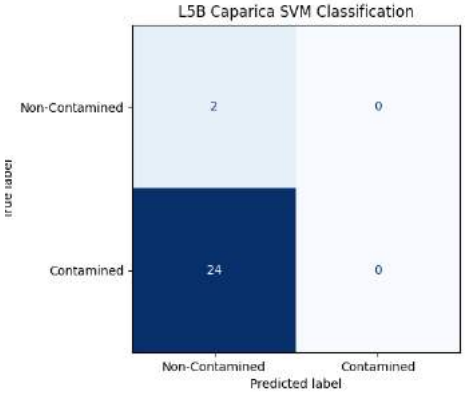


(a) L5b-D.

(b) L5b-DU.

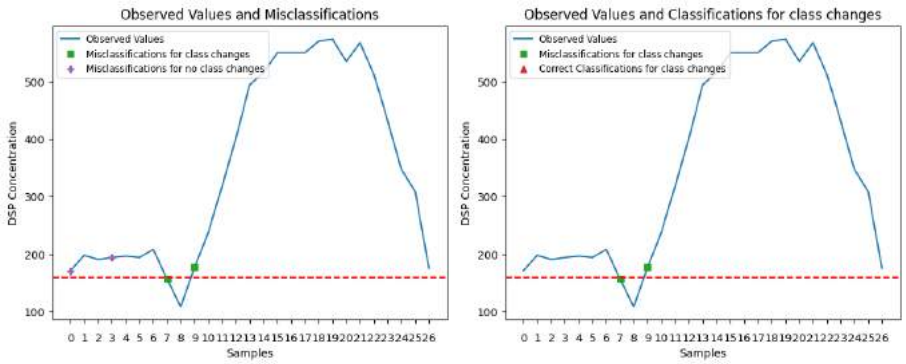


(c) L5b-DM.



(d) L5b-DMU.

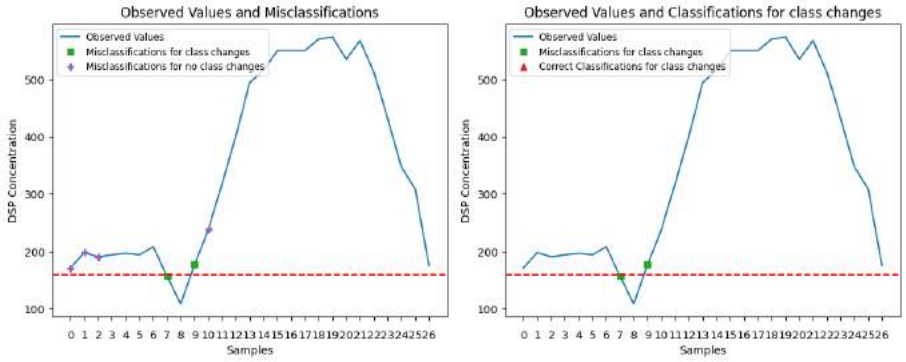
Figure C.59: SVM Classification Confusion Matrices for L5b Caparica.



(a) Missing distribution.

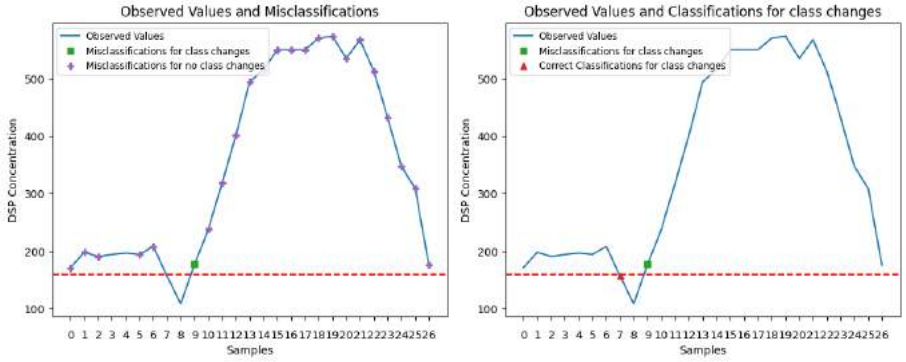
(b) Missing periods.

Figure C.60: SVM L5b-UP-D class changes and misclassifications.



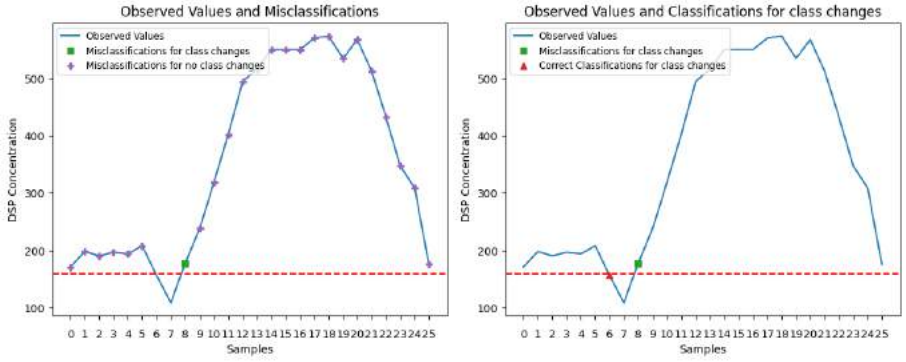
(a) Missing distribution. (b) Missing periods.

Figure C.61: SVM L5b-UP-DU class changes and misclassifications.



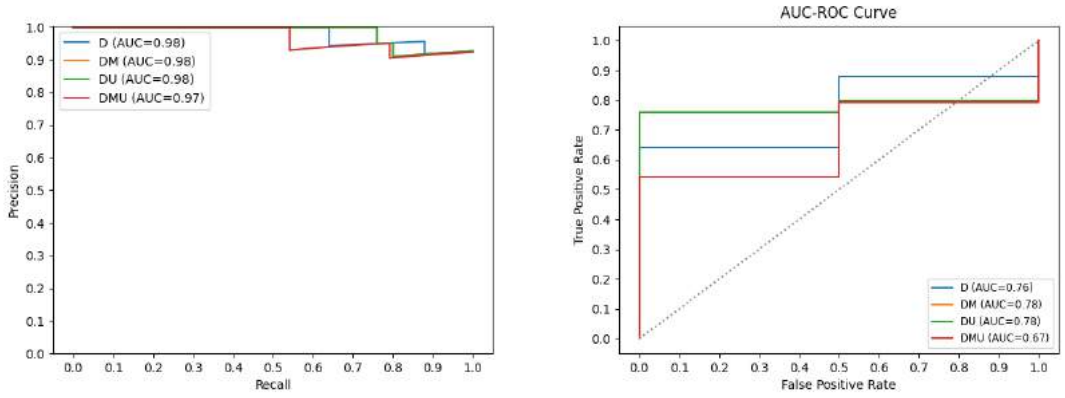
(a) Missing distribution. (b) Missing periods.

Figure C.62: SVM L5b-UP-DM class changes and misclassifications.



(a) Missing distribution. (b) Missing periods.

Figure C.63: SVM L5b-UP-DMU class changes and misclassifications.



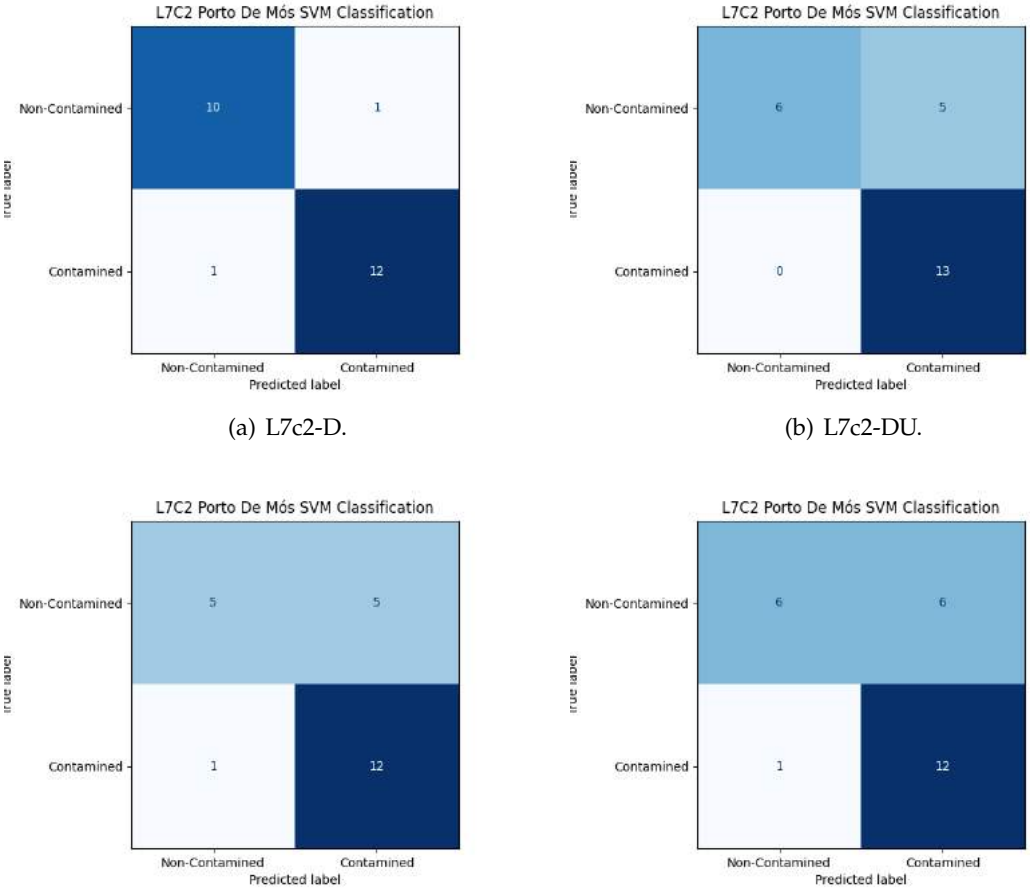
(a) PR-REC. (b) AUC-ROC.

Figure C.64: L5b Caparica SVM Classification models curves.

Table C.11: SVM Upwelling Classification metrics for L5b Caparica

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	Balanced Accuracy	0.9689	**0.9969**	0.9938	0.7900
	Average Precision	0.9983	**1.0000**	0.9998	0.9968
	F1 Macro	0.9597	**0.9953**	0.9907	0.7270
	Recall	0.9506	**0.9938**	0.9877	0.9133
Validation	Balanced Accuracy	**0.9246**	0.8710	0.8234	0.7813
	Average Precision	0.8952	**0.9076**	0.8259	0.8233
	F1 Macro	**0.9231**	0.8560	0.8346	0.7835
	Recall	**0.9444**	0.9167	0.6944	0.7186
Test	Balanced Accuracy	**0.6900**	0.6500	0.5400	0.5000
	Average Precision	0.9790	**0.9820**	**0.9820**	0.9682
	F1 Macro	**0.6250**	0.5598	0.1482	0.0714
	Recall	**0.8800**	0.8000	0.0800	0.0000

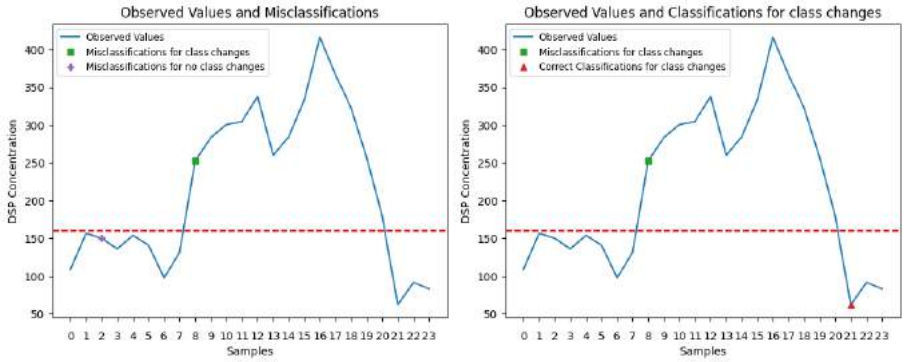
C.1.2.6 SVM Classification Upwelling L7c2 Porto de Mós



(a) L7c2-D. (b) L7c2-DU.

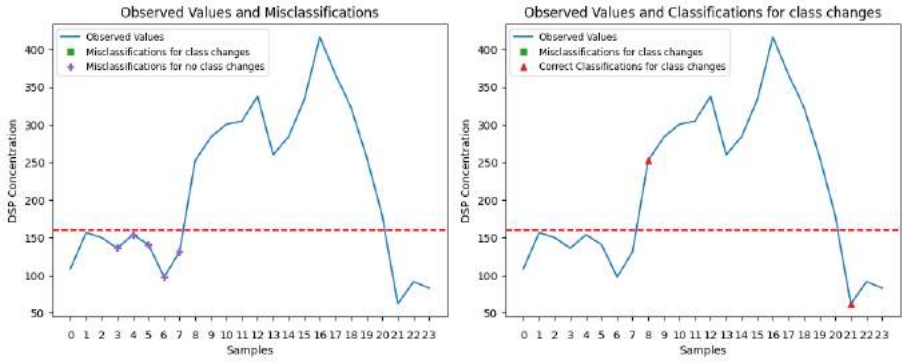
(c) L7c2-DM. (d) L7c2-DMU.

Figure C.65: SVM Classification Confusion Matrices for L7c2 Porto de Mós.



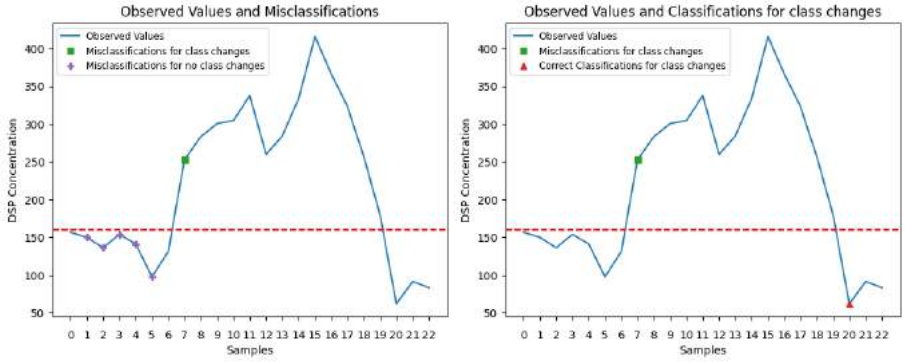
(a) Missing distribution. (b) Missing periods.

Figure C.66: SVM L7c2-UP-D class changes and misclassifications.



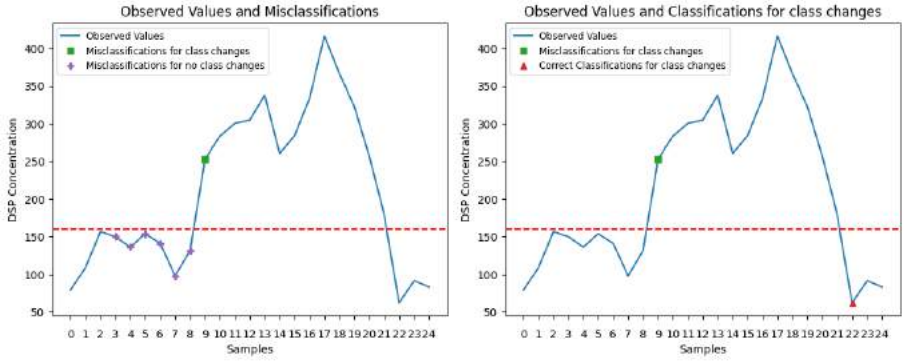
(a) Missing distribution. (b) Missing periods.

Figure C.67: SVM L7c2-UP-DU class changes and misclassifications.



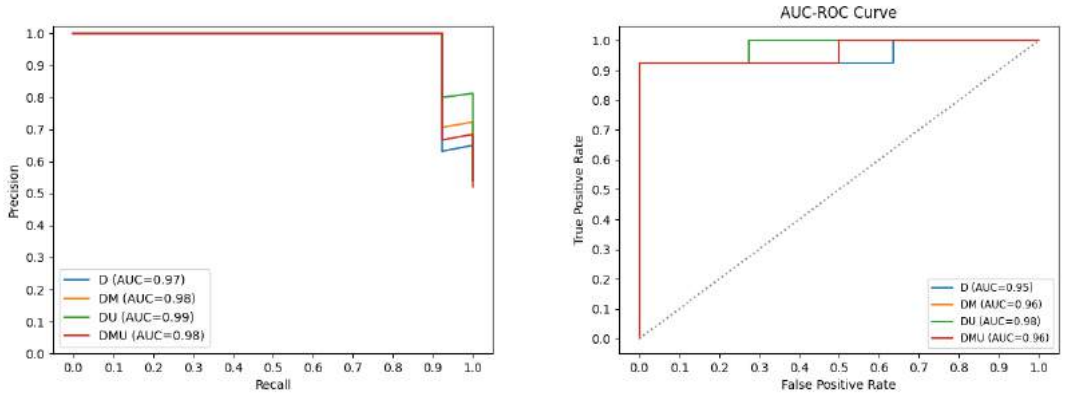
(a) Missing distribution. (b) Missing periods.

Figure C.68: SVM L7c2-UP-DM class changes and misclassifications.



(a) Missing distribution. (b) Missing periods.

Figure C.69: SVM L7c2-UP-DMU class changes and misclassifications.



(a) PR-REC. (b) AUC-ROC.

Figure C.70: L7c2 Porto de Mós SVM Classification models curves.

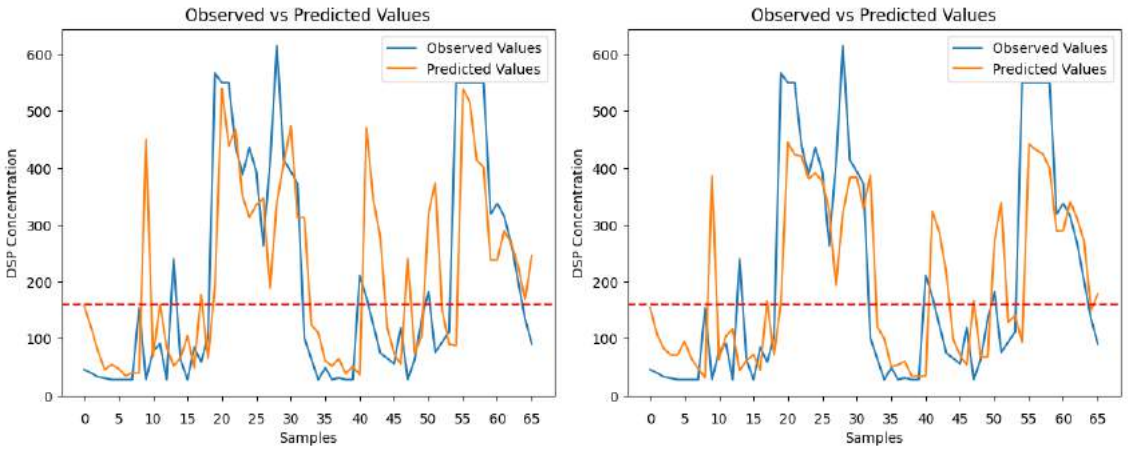
Table C.12: SVM Upwelling Classification metrics for L7c2 Porto de Mós

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	Balanced Accuracy	0.9136	0.9620	**1.000**	0.9748
	Average Precision	0.9868	0.9981	**1.000**	**1.0000**
	F1 Macro	0.9150	0.9614	**1.000**	0.9759
	Recall	0.8796	0.9460	**1.000**	0.9495
Validation	Balanced Accuracy	**0.9815**	0.9172	0.9310	0.8925
	Average Precision	**0.9986**	0.9264	0.8712	0.9705
	F1 Macro	**0.9848**	0.9143	0.9340	0.9080
	Recall	**1.000**	0.8889	0.8796	0.8458
Test	Balanced Accuracy	**0.9161**	0.7727	0.7115	0.7115
	Average Precision	0.9731	**0.9856**	0.9786	0.9757
	F1 Macro	**0.9161**	0.7723	0.7125	0.7029
	Recall	0.9231	**1.000**	0.9231	0.9231

C.2 Regression

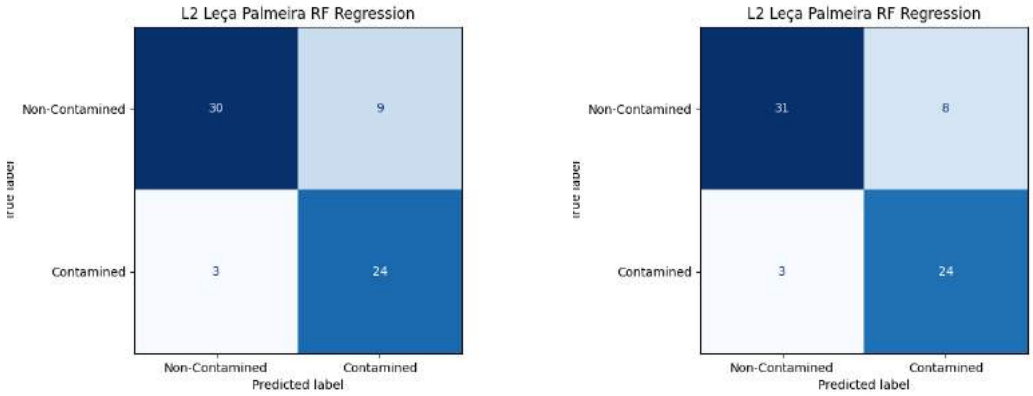
C.2.1 Random Forest

C.2.1.1 RF Regression L2 Leça da Palmeira



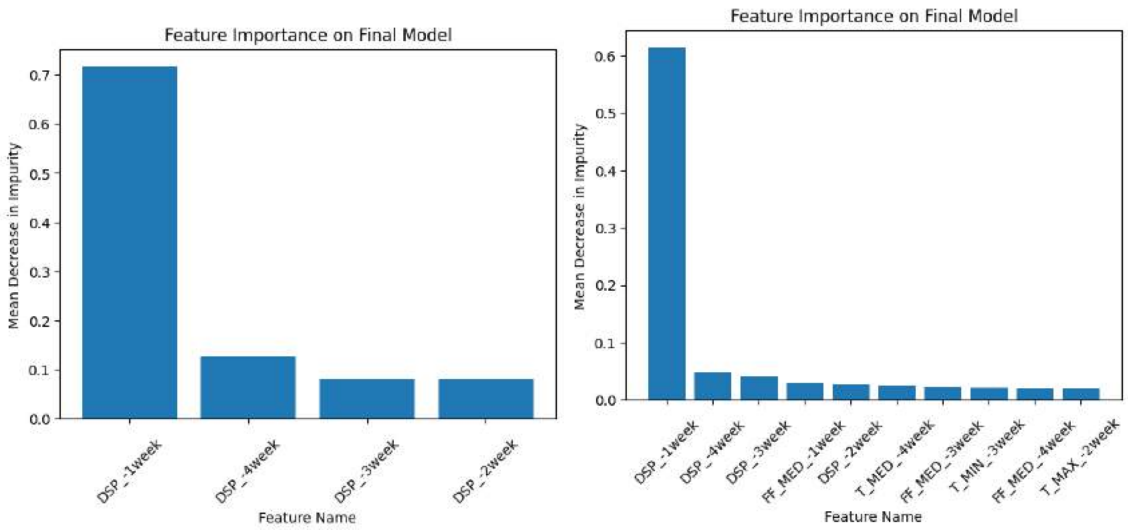
(a) l2-D. (b) L2-DM.

Figure C.71: RF Regression DSP predictions in L2 Leça da Palmeira.



(a) L2-D. (b) L2-DM.

Figure C.72: RF Regression Confusion Matrices for L2 Leça da Palmeira.



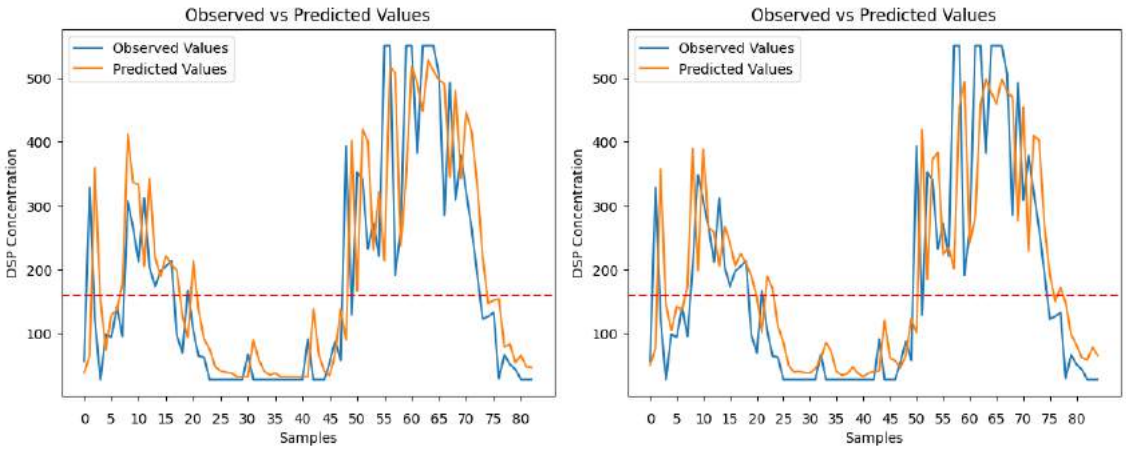
(a) L2-D. (b) L2-DM.

Figure C.73: L2 Leça Palmeira RF Regression feature importance.

Table C.13: RF Regression metrics for L2 Leça da Palmeira

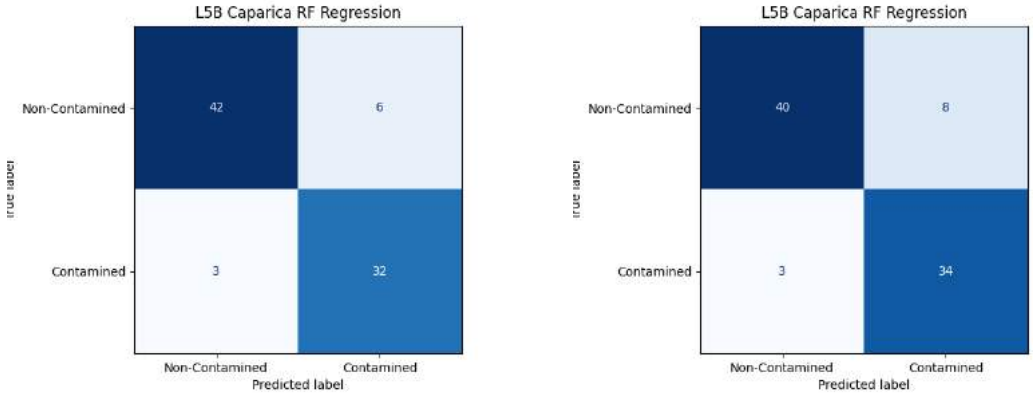
Set	Metric	Dataset	
		D	DM
Train	MSE	11756.54	**8618.24**
	RMSE	107.85	**92.33**
	MAE	61.56	**50.56**
	R2	0.66	**0.75**
Validation	MSE	**15977.92**	16340.31
	RMSE	**123.06**	125.79
	MAE	**79.48**	89.65
	R2	**0.48**	0.45
Test	MSE	20060.72	**17542.76**
	RMSE	141.64	**132.35**
	MAE	96.69	**90.30**
	R2	0.43	**0.50**

C.2.1.2 RF Regression L5b Caparica



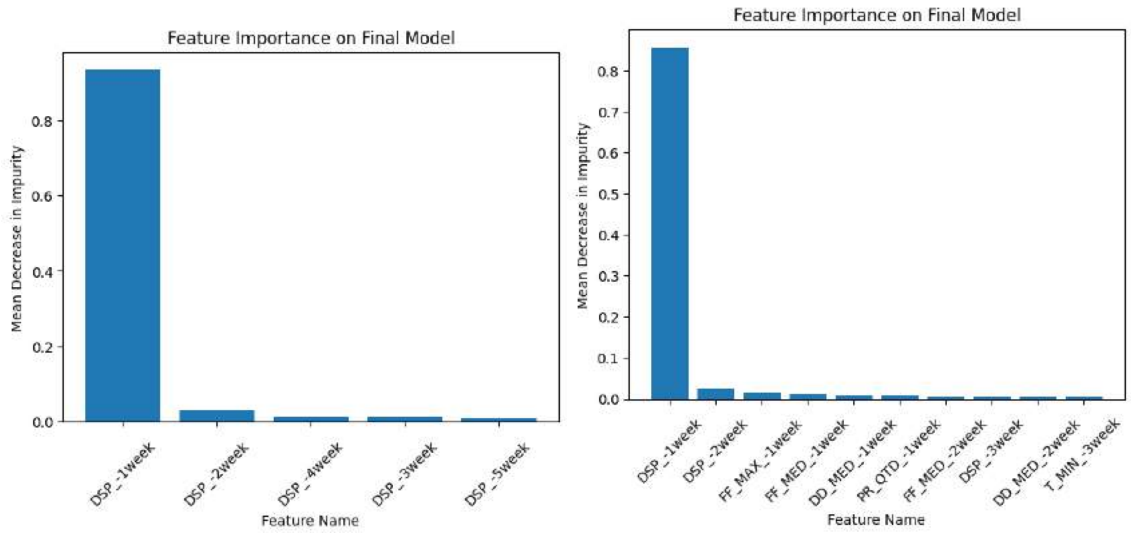
(a) L5b-D. (b) L5b-DM.

Figure C.74: RF Regression predictions for DSP in L5b Caparica.



(a) L5b-D. (b) L5b-DM.

Figure C.75: RF Regression Confusion Matrices for L5b Caparica.



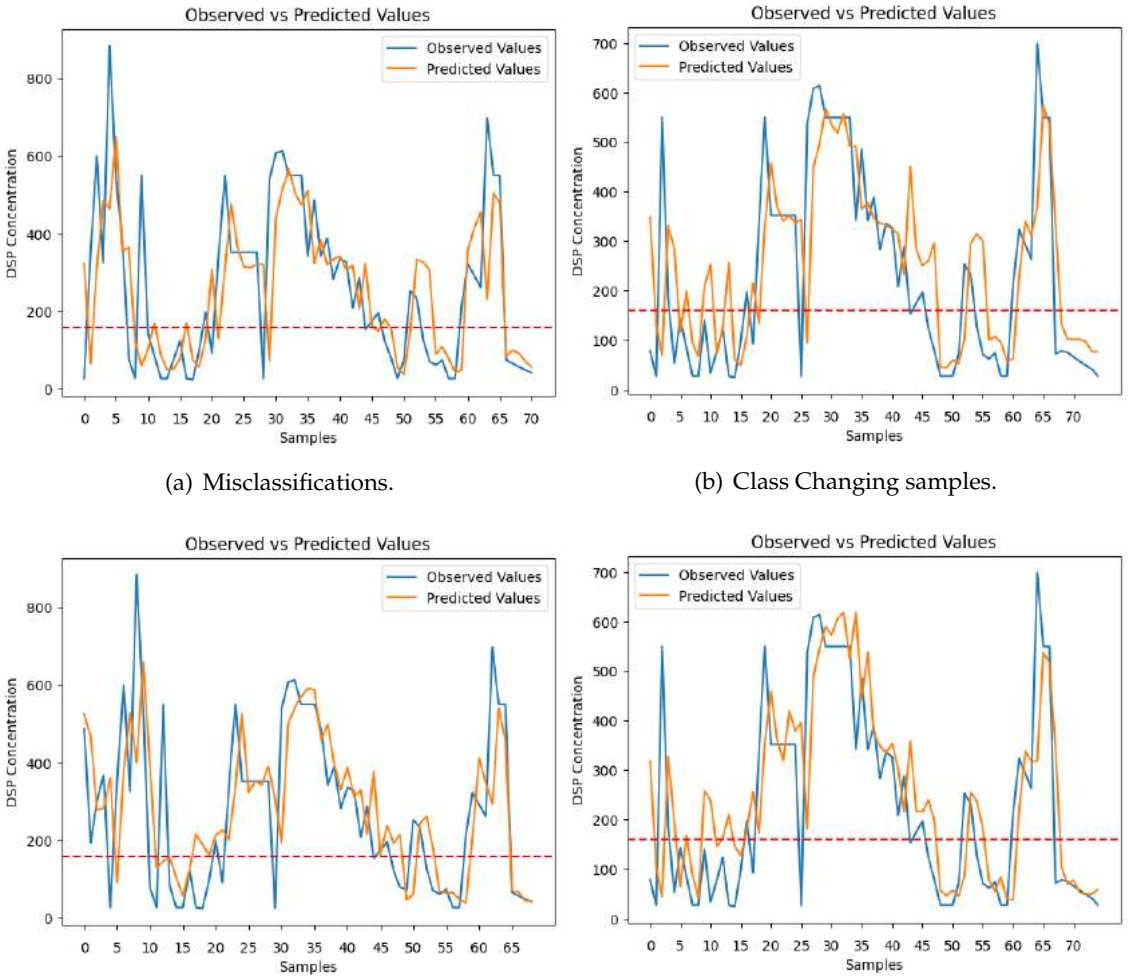
(a) L5b-D. (b) L5b-DM.

Figure C.76: L5b Caparica RF Regression feature importance.

Table C.14: RF Classification metrics for L5b-Caparica

Set	Metric	Dataset	
		D	DM
Train	MSE	20710.43	**18897.51**
	RMSE	143.18	**136.74**
	MAE	90.66	**86.16**
	R2	0.74	**0.76**
Validation	MSE	**22260.88**	22295.70
	RMSE	**141.65**	141.82
	MAE	**96.34**	97.94
	R2	**0.55**	0.54
Test	MSE	**11986.19**	12189.52
	RMSE	**109.48**	110.41
	MAE	**75.05**	76.71
	R2	**0.56**	0.55

C.2.1.3 RF Regression RIAV1 Triângulo



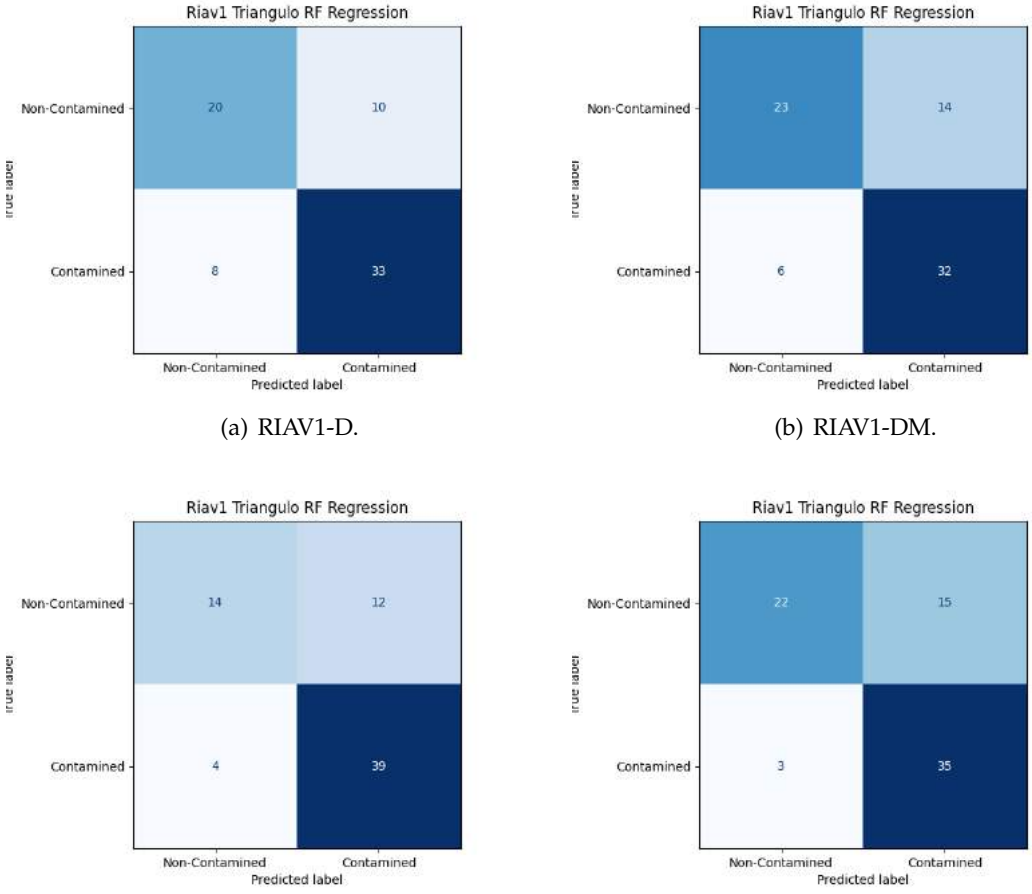
(a) Misclassifications.

(b) Class Changing samples.

(c) Misclassifications.

(d) Class Changing samples.

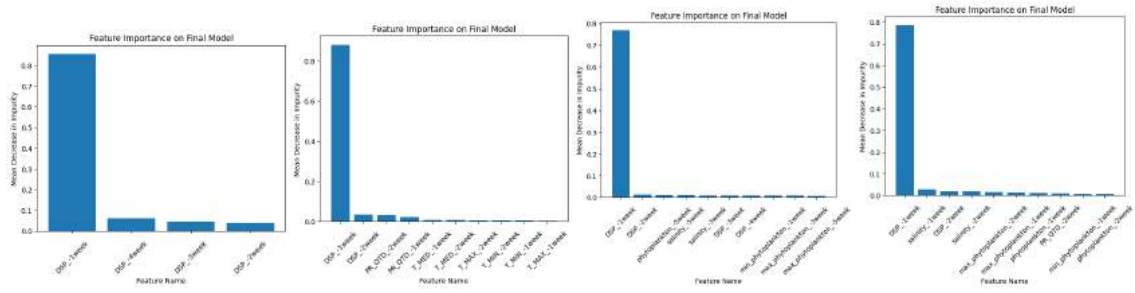
Figure C.77: RF Regression Predictions for RIAV1 Triângulo.



(a) RIAV1-D. (b) RIAV1-DM.

(c) RIAV1-DH. (d) RIAV1-DHM.

Figure C.78: RF Regression Confusion Matrices for RIAV1 Triângulo.



(a) RIAV1-D. (b) RIAV1-DM. (c) RIAV1-DH. (d) RIAV1-DHM.

Figure C.79: RIAV1 RF Regression feature importance.

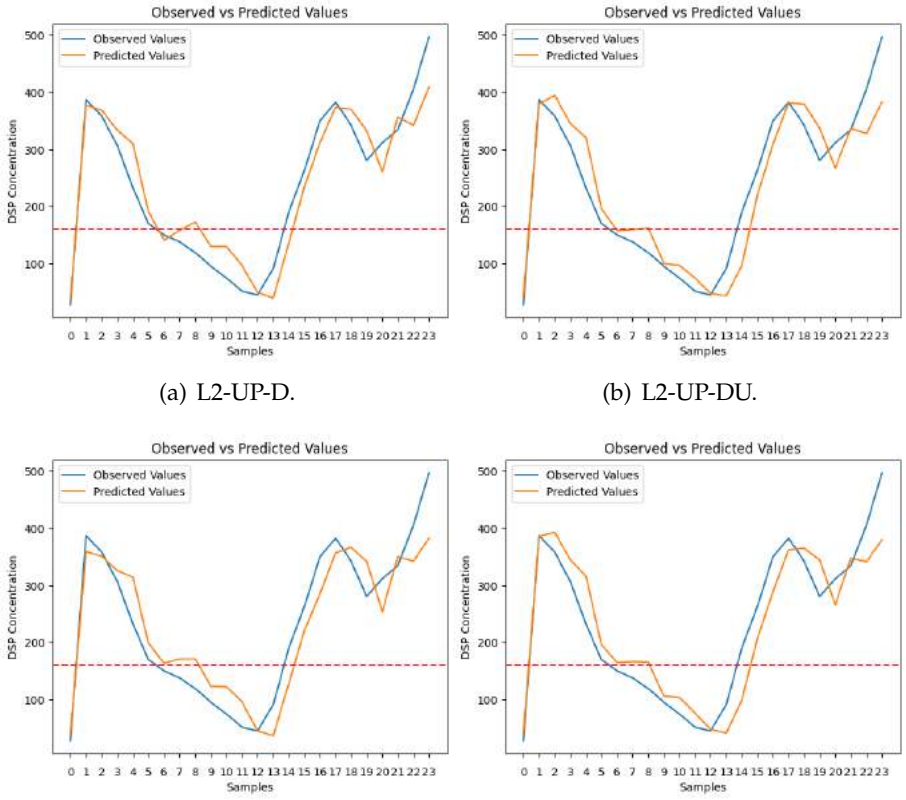
Table C.15: RF Regression metrics for RIAV1 Triângulo

Set	Metric	Dataset			
		D	DM	DH	DMH
Train	MSE	15621.85	20336.57	**13088.98**	16343.21
	RMSE	124.09	141.94	**113.28**	127.11
	MAE	76.22	90.04	**69.49**	80.52
	R2	0.73	0.66	**0.78**	0.72
Validation	MSE	**20527.64**	26200.79	24567.05	27374.75
	RMSE	**141.08**	157.69	152.04	161.07
	MAE	**103.62**	115.49	111.96	119.39
	R2	**0.50**	0.47	0.43	0.45
Test	MSE	26966.74	20777.29	24867.07	**18995.53**
	RMSE	164.22	144.14	157.69	**137.82**
	MAE	115.82	100.87	111.10	**94.49**
	R2	0.37	0.45	0.41	**0.49**

C.2.1.4 RF Regression Upwelling L2 Leça da Palmeira

Table C.16: RF Upwelling Regression metrics for L2 Leça da Palmeira

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	MSE	7011.57	6160.16	6613.37	**5987.56**
	RMSE	80.95	75.01	78.28	**73.91**
	MAE	53.09	48.12	48.38	**46.25**
	R2	0.81	0.69	0.82	**0.84**
Validation	MSE	**4445.17**	5437.51	5655.10	5991.81
	RMSE	**65.38**	73.22	74.75	77.10
	MAE	**57.64**	63.78	66.44	67.14
	R2	**0.77**	0.69	0.69	0.66
Test	MSE	**1808.58**	2299.95	2354.00	2426.59
	RMSE	**42.53**	47.96	48.52	49.26
	MAE	**35.84**	37.24	40.92	39.99
	R2	**0.90**	0.87	0.87	0.86



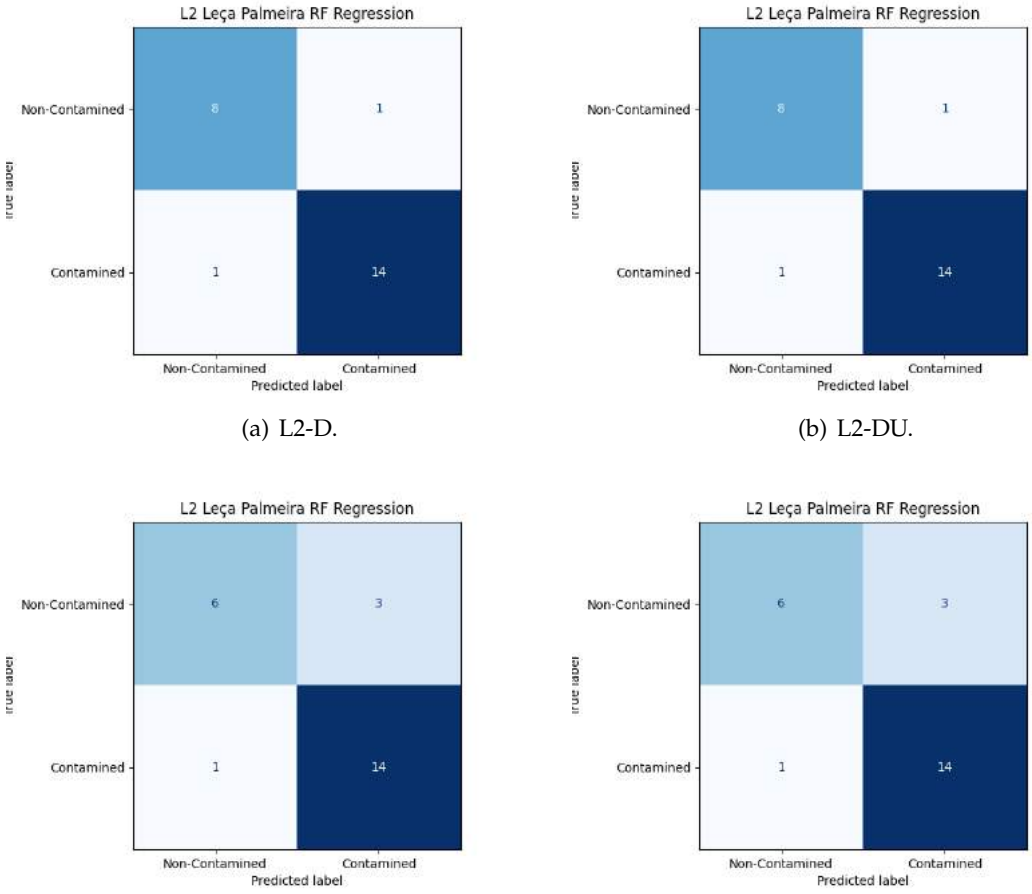
(a) L2-UP-D.

(b) L2-UP-DU.

(c) L2-DU-DM.

(d) L2-UP-DMU.

Figure C.80: RF L2 Leça da Palmeira DSP predictions.



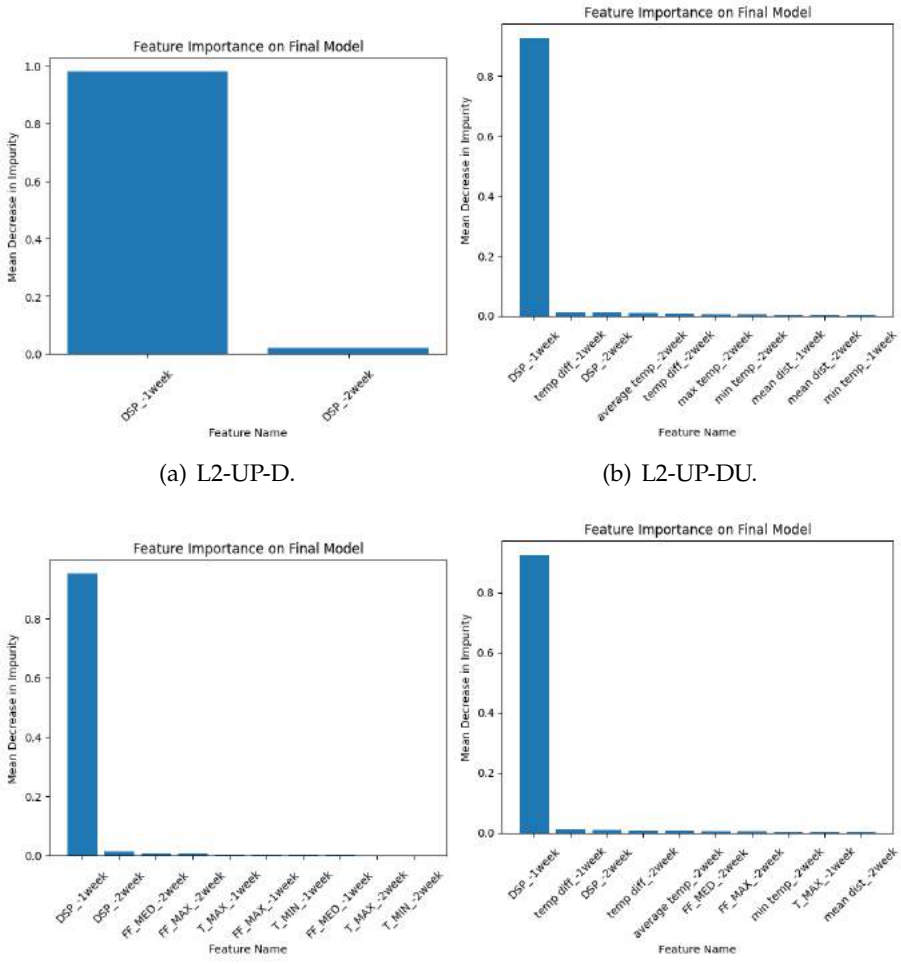
(a) L2-D.

(b) L2-DU.

(c) L2-DM.

(d) L2-DMU.

Figure C.81: RF Regression Confusion Matrices for L2 Leça da Palmeira.

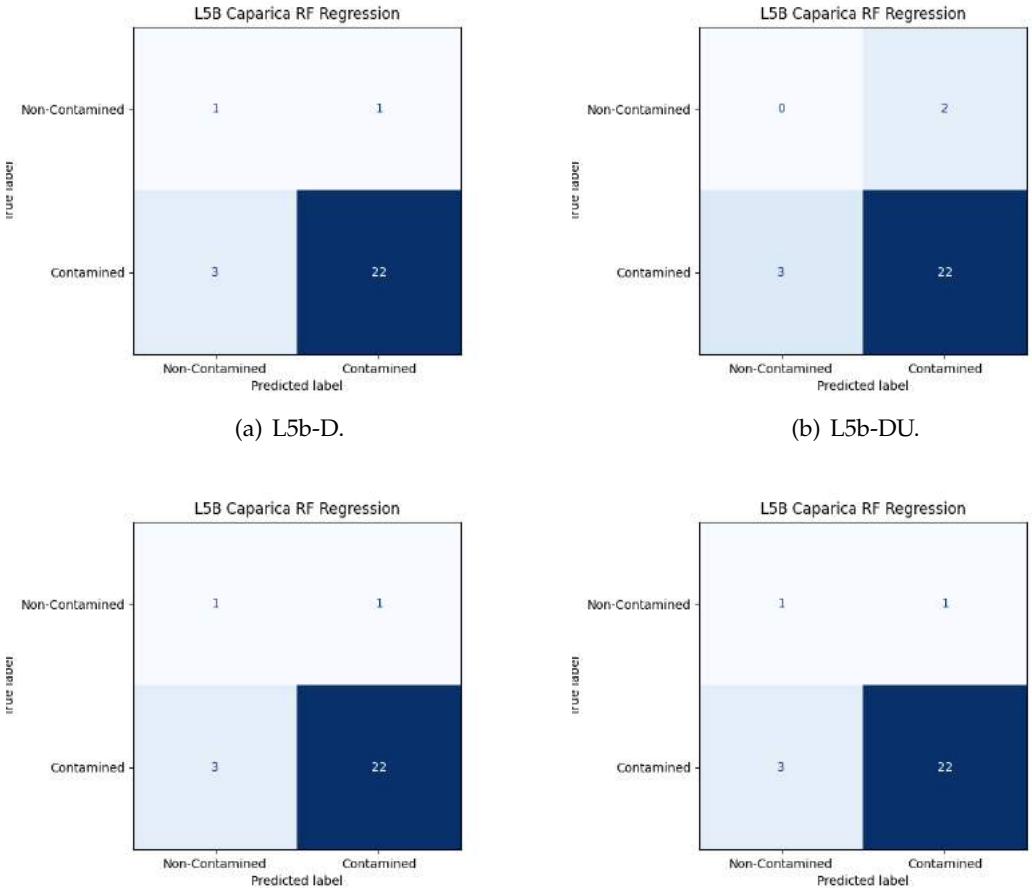


(a) L2-UP-D. (b) L2-UP-DU.

(c) L2-DU-DM. (d) L2-UP-DMU.

Figure C.82: RF Regression L2 Leça da Palmeira Feature Importance.

C.2.1.5 RF Regression Upwelling L5b Caparica



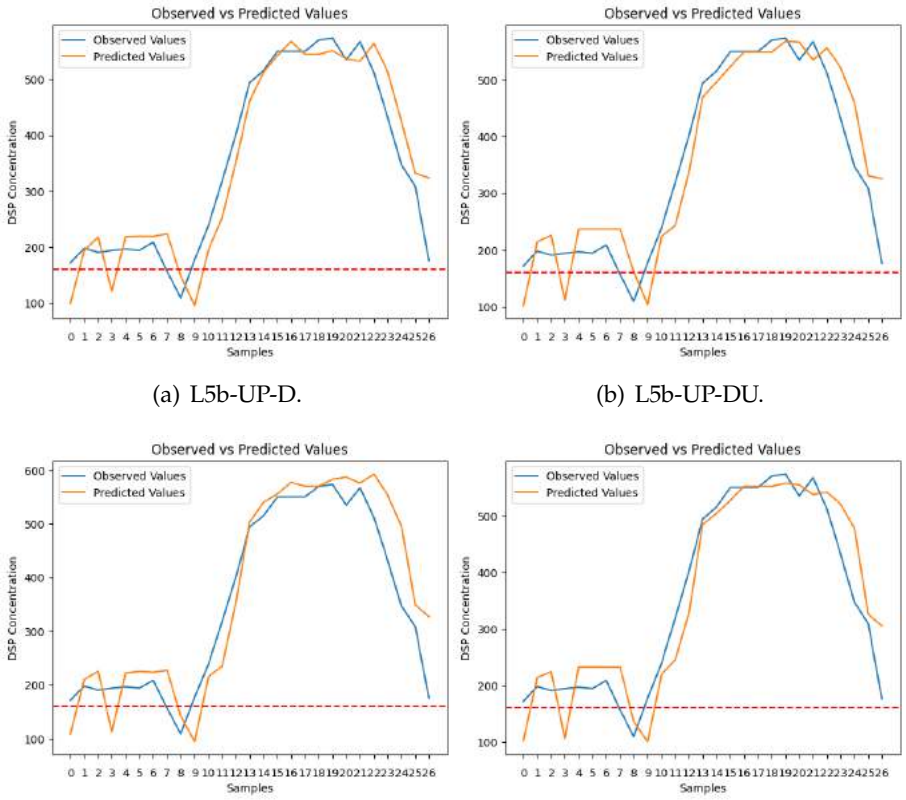
(a) L5b-D.

(b) L5b-DU.

(c) L5b-DM.

(d) L5b-DMU.

Figure C.83: RF Regression Confusion Matrices for L5b Caparica.



(a) L5b-UP-D.

(b) L5b-UP-DU.

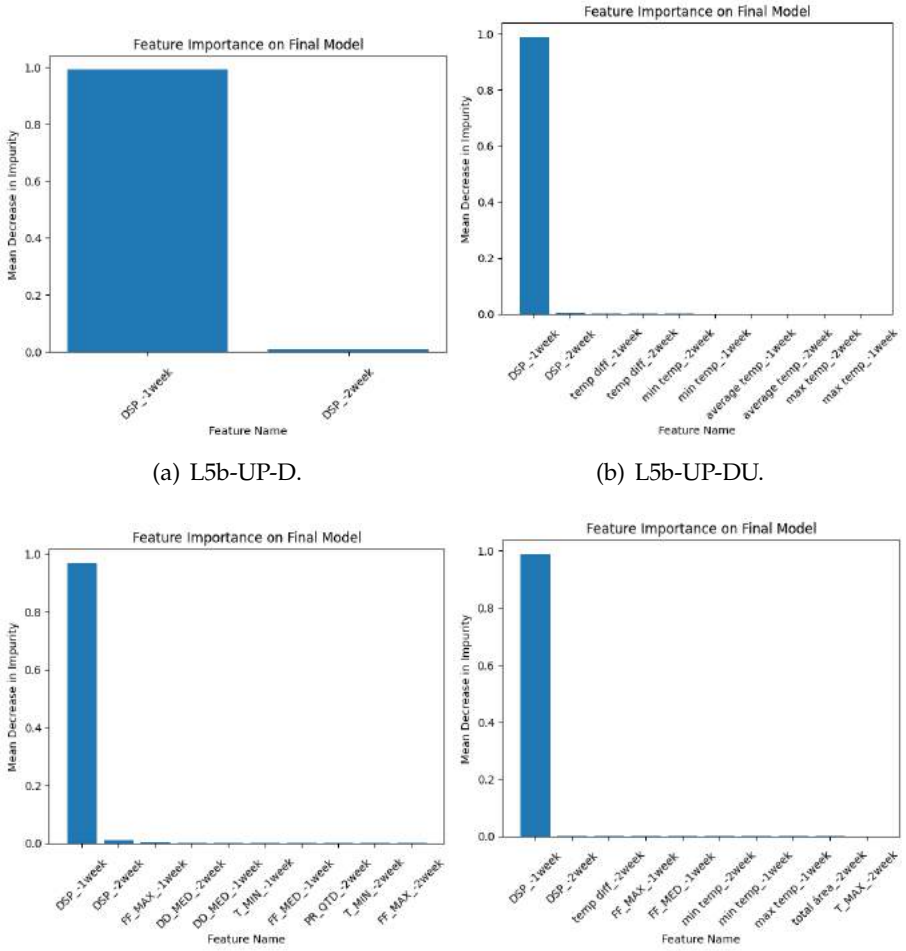
(c) L5b-DU-DM.

(d) L5b-UP-DMU.

Figure C.84: RF L5b Caparica DSP predictions.

Table C.17: RF Upwelling Regression metrics for L5b Caparica

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	MSE	2555.79	4506.73	**2020.77**	4281.00
	RMSE	49.38	66.38	**44.11**	64.73
	MAE	35.99	53.16	**32.03**	51.81
	R2	0.96	0.93	**0.97**	0.93
Validation	MSE	**21196.17**	27974.58	23654.33	30162.64
	RMSE	**119.90**	138.03	125.51	141.76
	MAE	**95.22**	114.49	98.48	117.31
	R2	**0.40**	0.21	0.34	0.16
Test	MSE	**2790.75**	3400.91	4040.56	3192.19
	RMSE	**52.83**	58.32	63.57	56.50
	MAE	**41.19**	46.55	48.29	43.59
	R2	**0.90**	0.87	0.85	0.88



(a) L5b-UP-D.

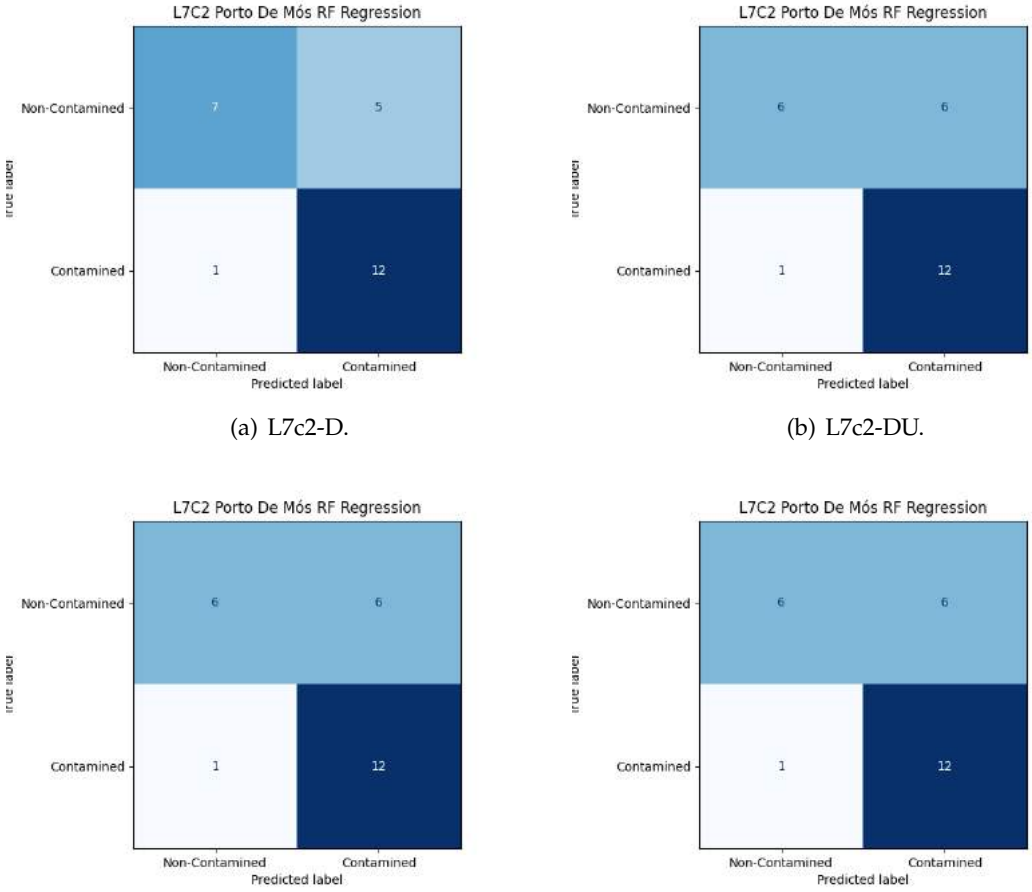
(b) L5b-UP-DU.

(c) L5b-DU-DM.

(d) L5b-UP-DMU.

Figure C.85: RF L5b Caparica Feature Importance.

C.2.1.6 RF Regression Upwelling L7c2 Porto de Mós



(a) L7c2-D.

(b) L7c2-DU.

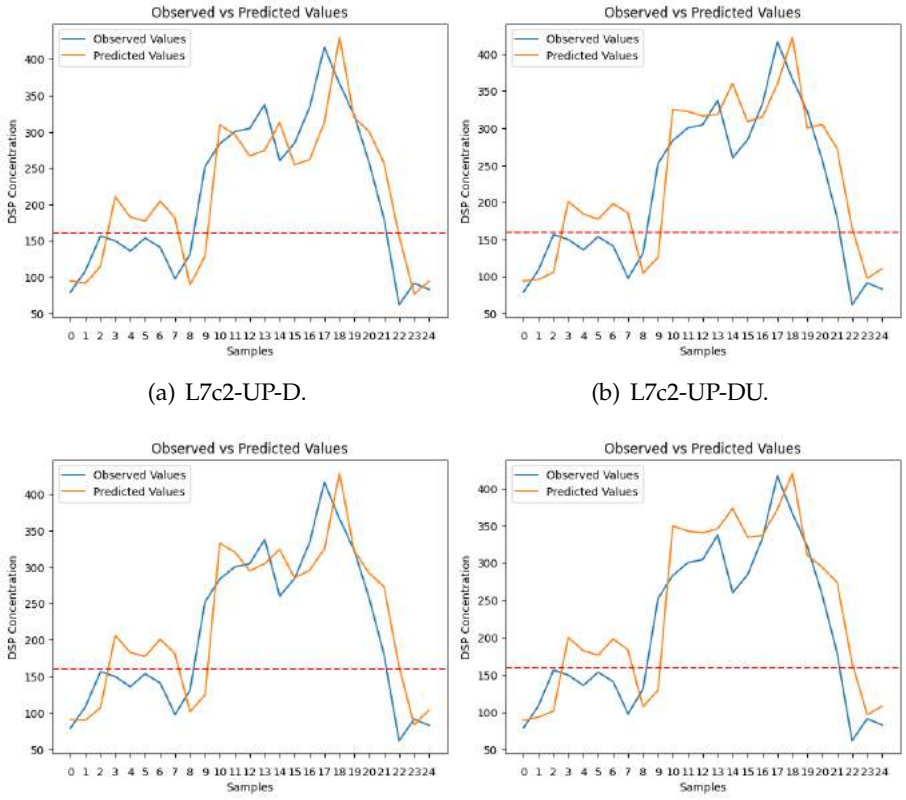
(c) L7c2-DM.

(d) L7c2-DMU.

Figure C.86: RF Regression Confusion Matrices for L7c2 Porto de Mós.

Table C.18: RF Upwelling Regression metrics for L7c2 Porto de Mós

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	MSE	2824.70	2070.35	1932.15	**1881.43**
	RMSE	52.69	44.97	43.50	**42.87**
	MAE	39.15	33.80	31.79	**31.44**
	R2	0.78	0.84	**0.85**	**0.85**
Validation	MSE	**3999.60**	4685.55	8046.36	8148.18
	RMSE	**59.04**	64.02	80.08	80.28
	MAE	**49.49**	54.30	64.78	65.35
	R2	**0.78**	0.73	0.59	0.58
Test	MSE	3337.54	3180.79	**3173.02**	3387.10
	RMSE	57.77	56.40	**56.33**	58.20
	MAE	48.56	46.09	**45.44**	47.47
	R2	0.68	**0.70**	**0.70**	0.68



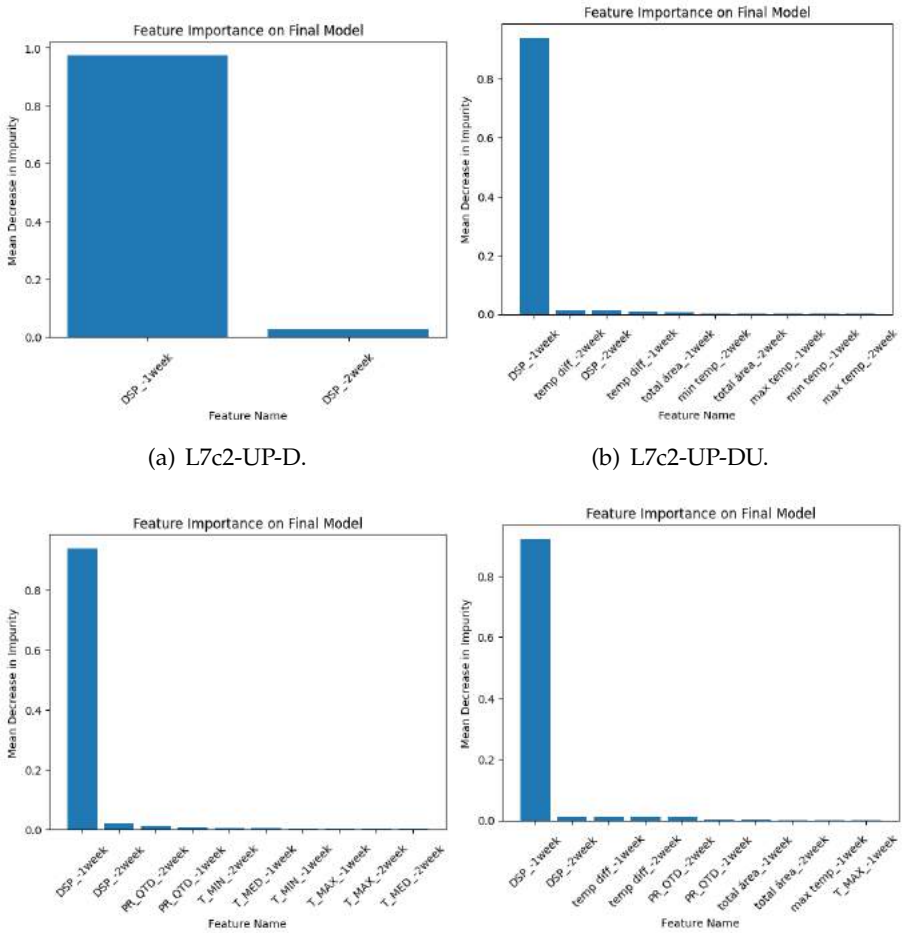
(a) L7c2-UP-D.

(b) L7c2-UP-DU.

(c) L7c2-DU-DM.

(d) L7c2-UP-DMU.

Figure C.87: RF L7c2 Porto de Mós DSP predictions.



(a) L7c2-UP-D.

(b) L7c2-UP-DU.

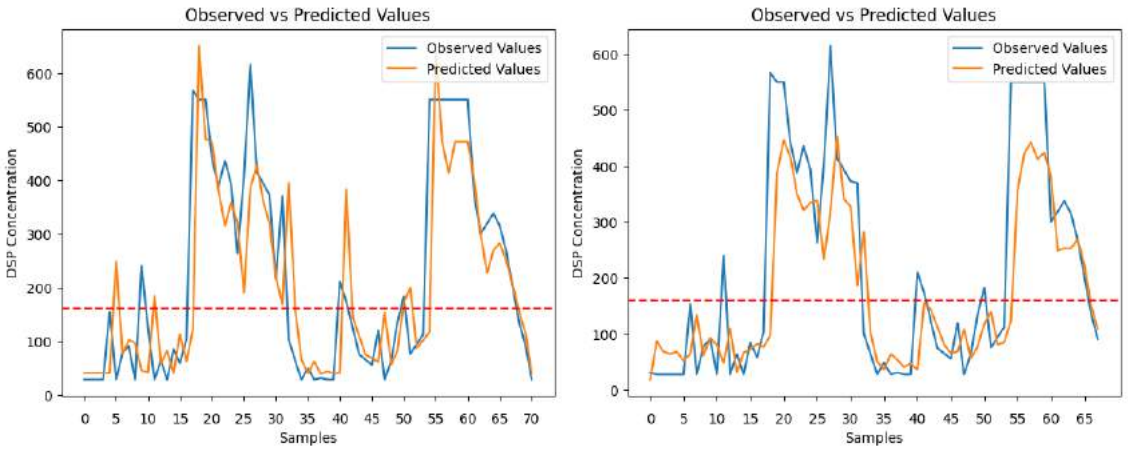
(c) L7c2-DU-DM.

(d) L7c2-UP-DMU.

Figure C.88: RF L7c2 Porto de Mós Feature Importance.

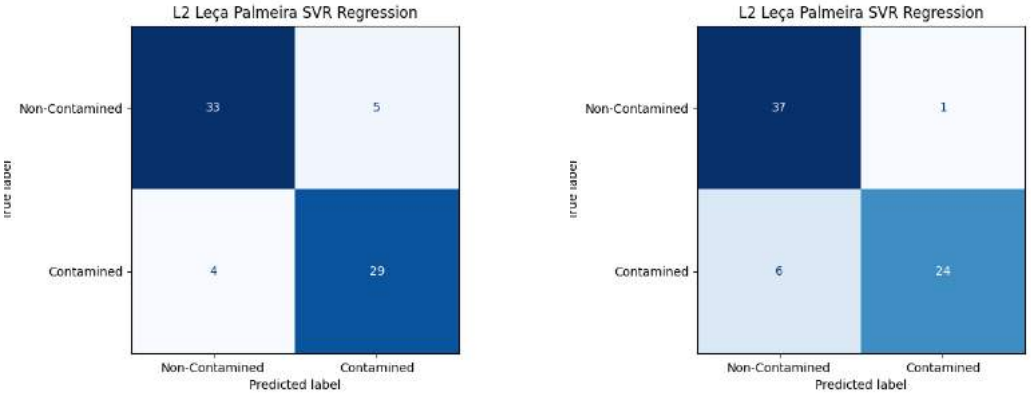
C.2.2 Support Vector Regression

C.2.2.1 SVR Regression L2 Leça da Palmeira



(a) l2-D. (b) L2-DM.

Figure C.89: SVR Regression DSP predictions in L2 Leça da Palmeira.



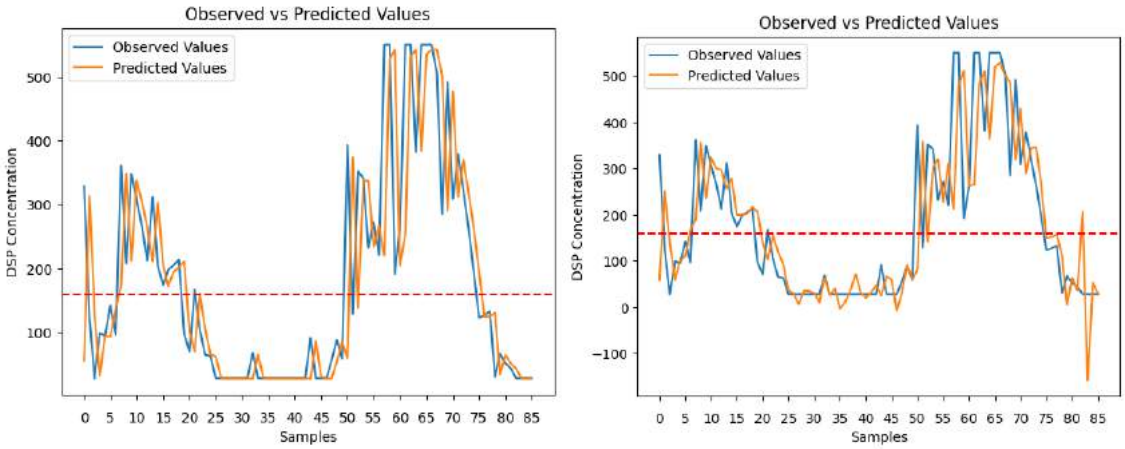
(a) L2-D. (b) L2-DM.

Figure C.90: SVR Regression Confusion Matrices for L2 Leça da Palmeira.

Table C.19: SVR Regression metrics for L2 Leça da Palmeira

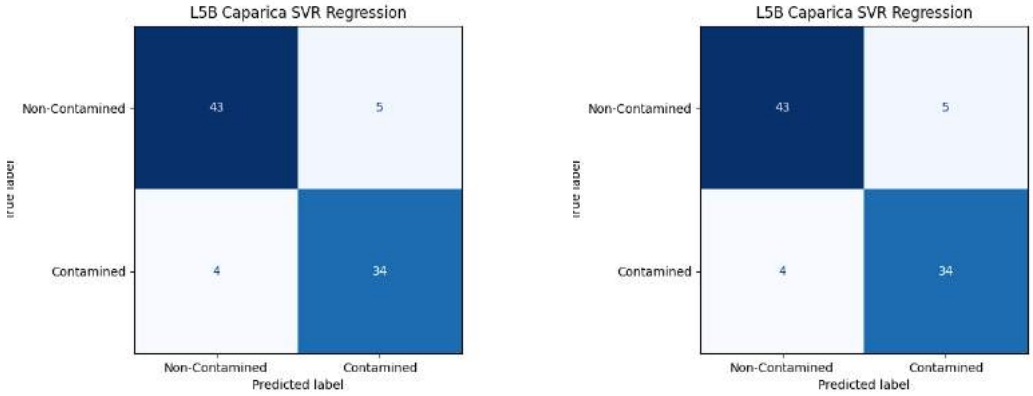
Set	Metric	Dataset	
		D	DM
Train	MSE	20345.70	**19404.35**
	RMSE	142.26	**139.13**
	MAE	66.74	**63.96**
	R2	0.44	**0.46**
Validation	MSE	**15449.62**	18337.93
	RMSE	**120.78**	133.17
	MAE	**73.38**	82.13
	R2	**0.52**	0.42
Test	MSE	13935.77	**13420.91**
	RMSE	118.05	**115.85**
	MAE	**75.37**	76.01
	R2	0.62	**0.63**

C.2.2.2 SVR Regression L5b Caparica



(a) L5b-D. (b) L5b-DM.

Figure C.91: SVR Regression predictions for DSP in L5b Caparica.



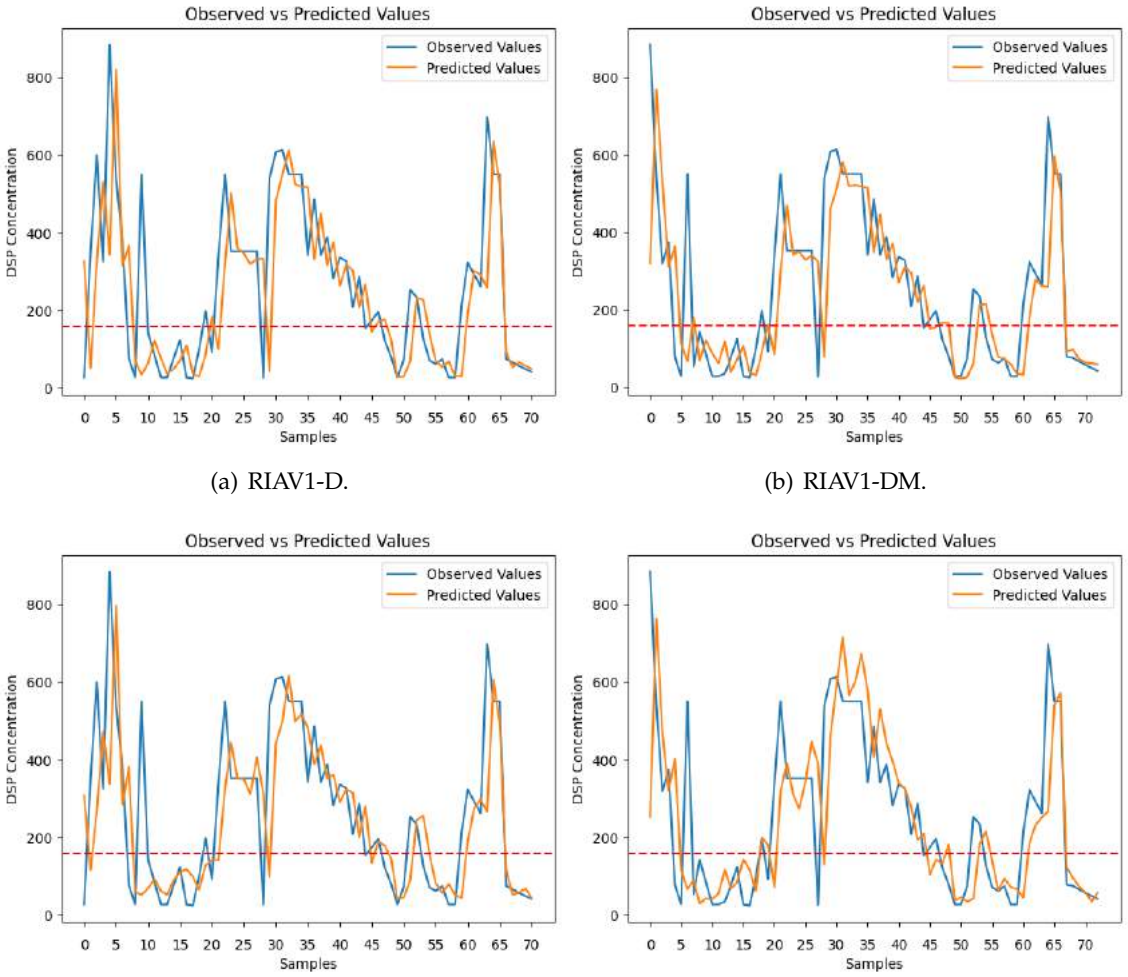
(a) L5b-D. (b) L5b-DM.

Figure C.92: SVR Regression Confusion Matrices for L5b Caparica.

Table C.20: SVR Classification metrics for L5b Caparica

Set	Metric	Dataset	
		D	DM
Train	MSE	26854.86	**24798.00**
	RMSE	162.99	**156.60**
	MAE	**82.74**	88.89
	R2	0.66	**0.68**
Validation	MSE	21342.41	**19748.45**
	RMSE	136.95	**132.96**
	MAE	**76.03**	82.47
	R2	0.59	**0.61**
Test	MSE	11987.50	**11525.96**
	RMSE	109.49	**107.36**
	MAE	**67.30**	70.25
	R2	0.56	**0.58**

C.2.2.3 SVR Regression RIAV1 Triângulo



(a) RIAV1-D.

(b) RIAV1-DM.

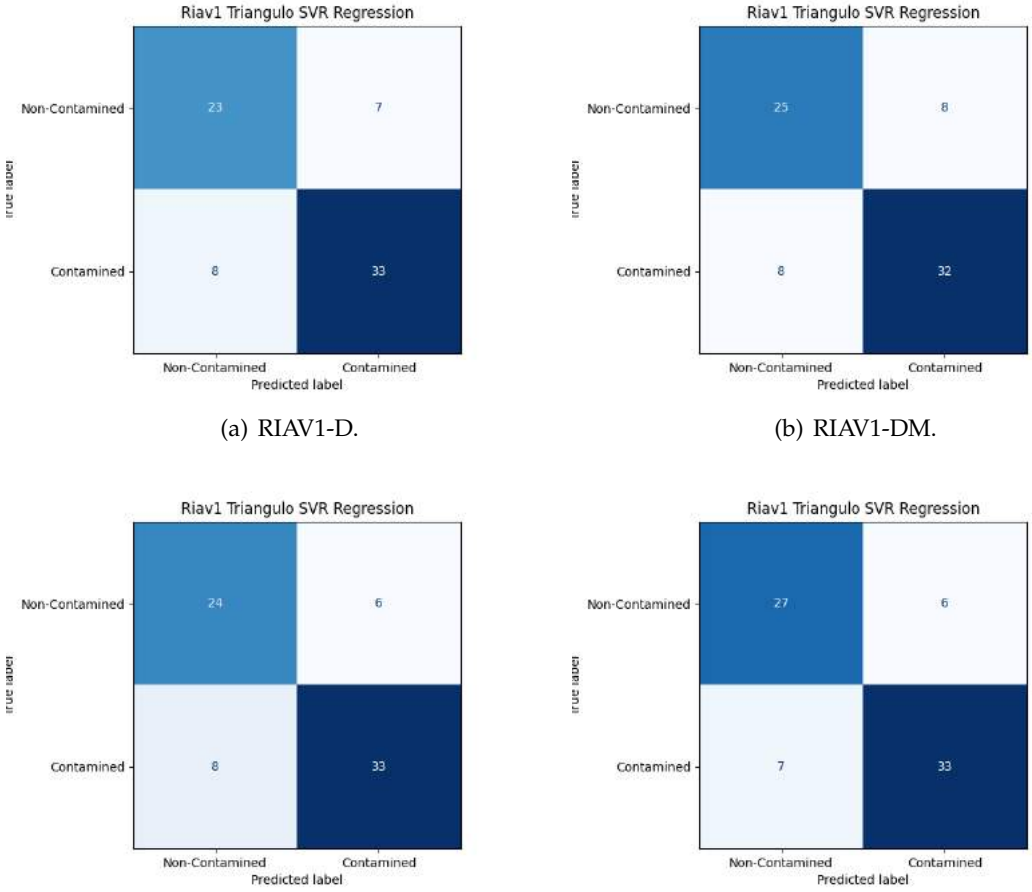
(c) RIAV1-DH.

(d) RIAV1-DMH.

Figure C.93: SVR Regression Predictions for RIAV1 Triângulo.

Table C.21: SVR Regression metrics for RIAV1 Triângulo

Set	Metric	Dataset			
		D	DM	DH	DMH
Train	MSE	26152.23	26194.85	**20262.06**	21707.67
	RMSE	160.87	161.27	**141.96**	146.96
	MAE	88.91	89.39	76.83	**75.84**
	R2	0.55	0.56	**0.65**	0.64
Validation	MSE	**20098.89**	23858.80	23606.27	28110.06
	RMSE	**138.16**	149.74	149.25	161.28
	MAE	**89.07**	104.13	107.04	109.53
	R2	**0.54**	0.46	0.45	0.38
Test	MSE	28068.54	**23310.23**	25732.85	25177.27
	RMSE	167.54	**152.68**	160.41	158.67
	MAE	109.53	**97.19**	105.37	101.08
	R2	0.34	**0.45**	0.40	0.41



(a) RIAV1-D.

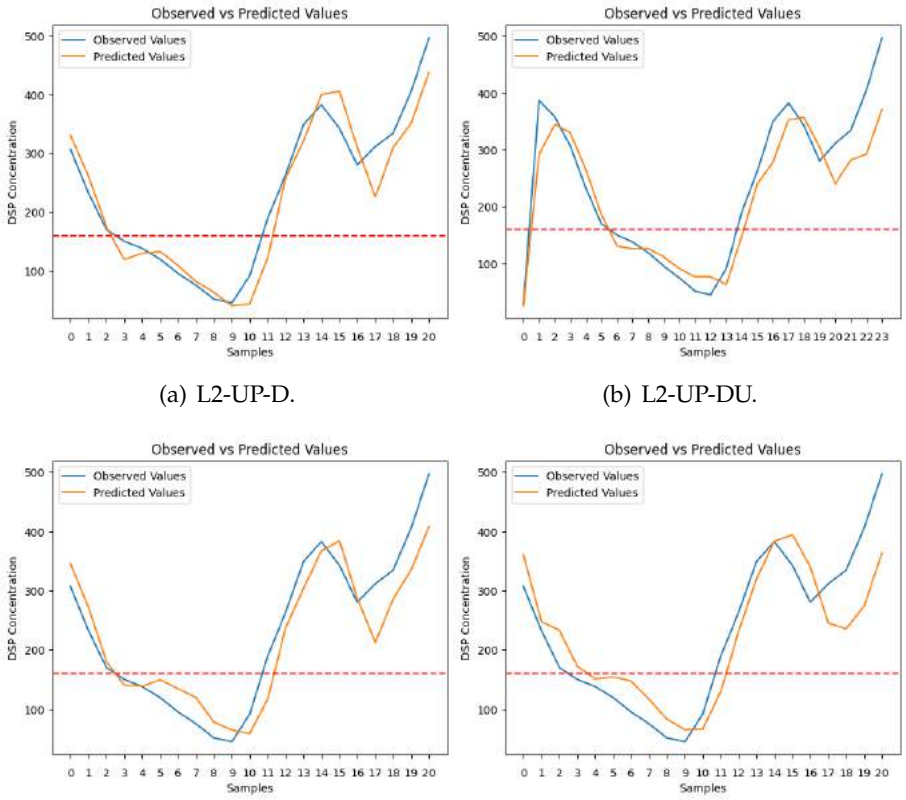
(b) RIAV1-DM.

(c) RIAV1-DH.

(d) RIAV1-DHM.

Figure C.94: SVR Regression Confusion Matrices for RIAV1 Triângulo.

C.2.2.4 SVR Regression Upwelling L2 Leça da Palmeira



(a) L2-UP-D.

(b) L2-UP-DU.

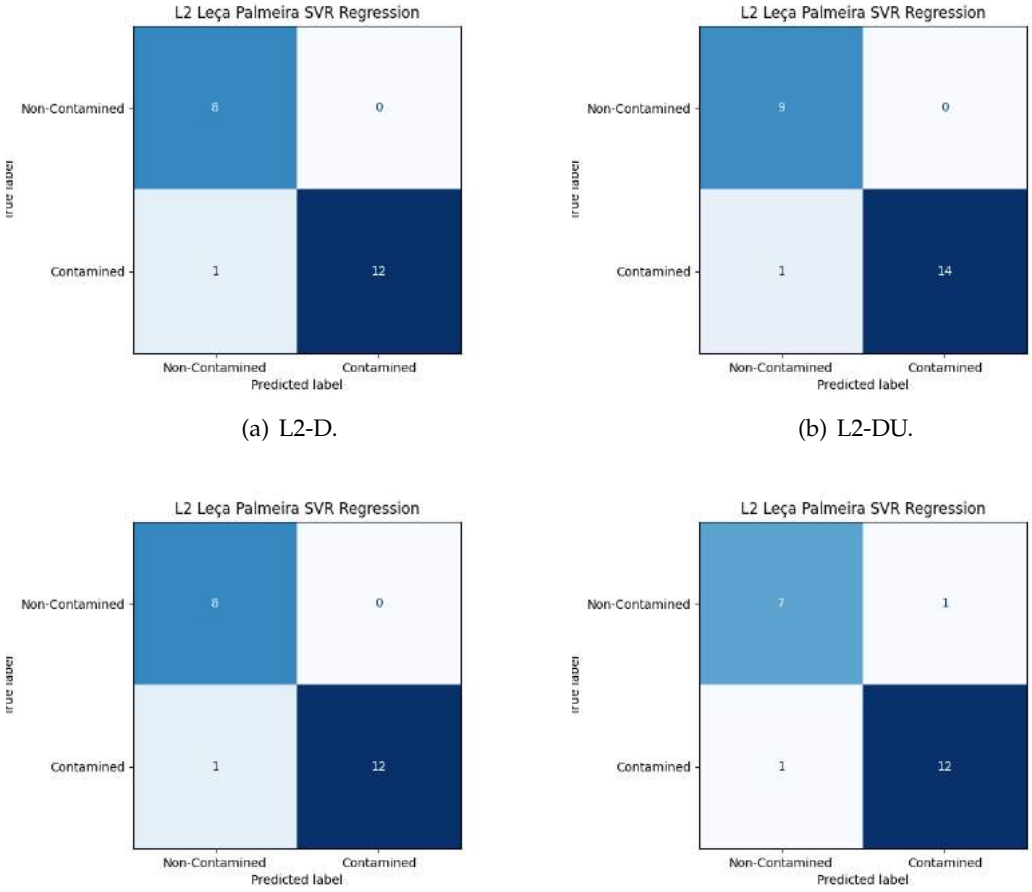
(c) L2-DU-DM.

(d) L2-UP-DMU.

Figure C.95: SVR L2 Leça da Palmeira DSP predictions.

Table C.22: SVR Upwelling Regression metrics for L2 Leça da Palmeira

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	MSE	1334.21	1183.64	905.01	**423.98**
	RMSE	36.51	34.37	29.60	**20.18**
	MAE	26.26	26.19	19.34	**13.85**
	R2	0.89	0.96	0.93	**0.97**
Validation	MSE	**1274.28**	5722.70	4363.01	8015.98
	RMSE	**34.86**	72.54	61.41	87.36
	MAE	**27.11**	58.41	48.25	73.30
	R2	**0.92**	0.71	0.60	0.31
Test	MSE	**1443.88**	2503.46	2143.80	3653.37
	RMSE	**38.00**	50.03	46.30	60.44
	MAE	**30.01**	37.72	38.44	49.39
	R2	**0.91**	0.86	0.87	0.77



(a) L2-D.

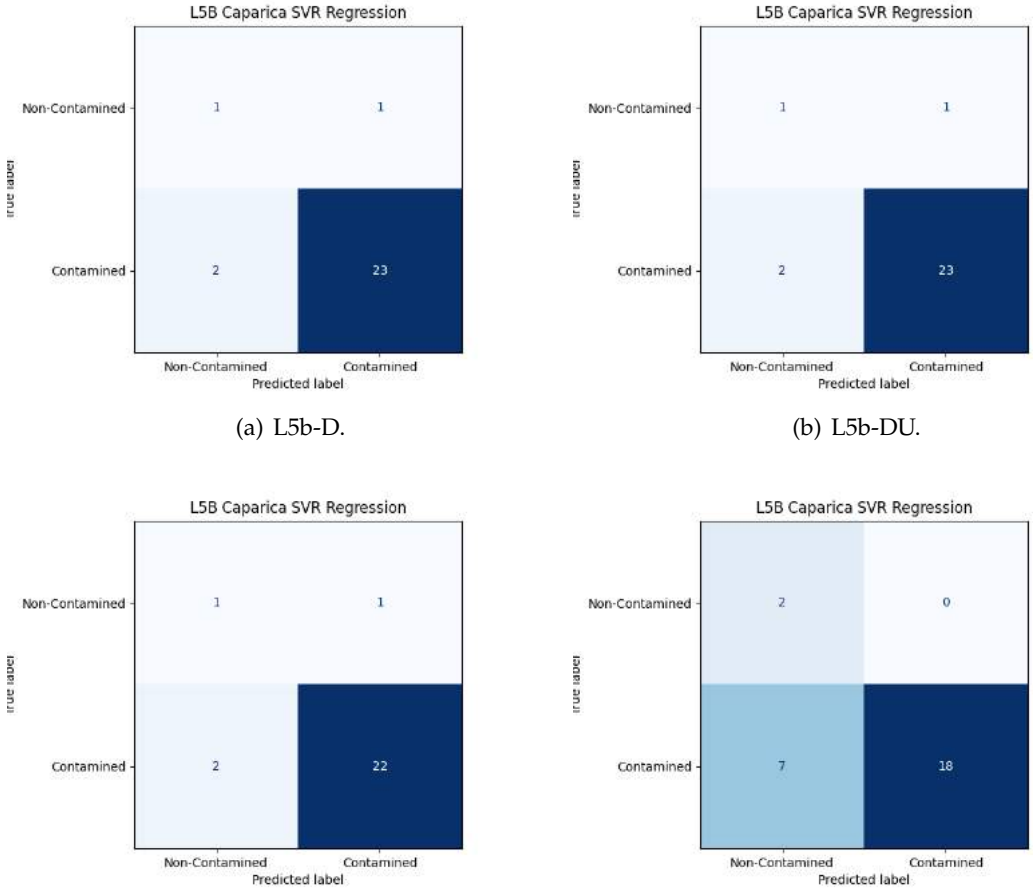
(b) L2-DU.

(c) L2-DM.

(d) L2-DMU.

Figure C.96: SVR Regression Confusion Matrices for L2 Leça da Palmeira.

C.2.2.5 SVR Regression Upwelling L5b Caparica



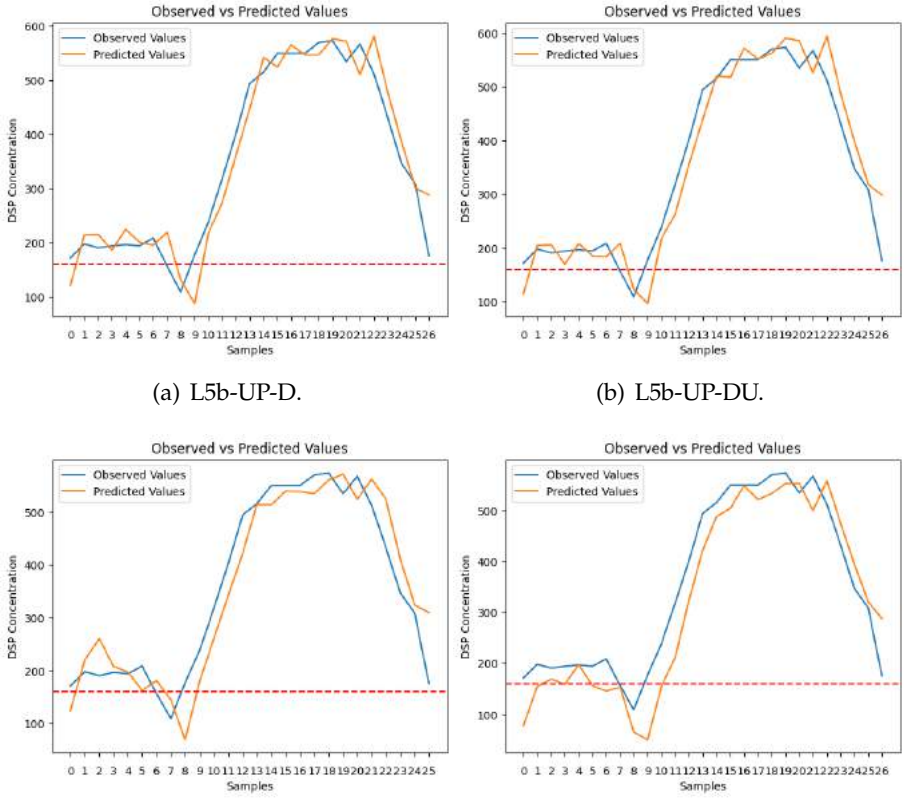
(a) L5b-D. (b) L5b-DU.

(c) L5b-DM. (d) L5b-DMU.

Figure C.97: SVR Regression Confusion Matrices for L5b Caparica.

Table C.23: SVR Upwelling Regression metrics for L5b Caparica

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	MSE	3509.89	**2451.34**	3635.21	2507.29
	RMSE	58.88	**48.71**	60.11	49.37
	MAE	41.56	**32.85**	40.81	33.75
	R2	0.94	**0.96**	0.94	0.96
Validation	MSE	**3566.97**	5821.04	7038.68	8926.43
	RMSE	**54.17**	66.65	73.89	90.66
	MAE	**40.74**	48.92	58.85	69.65
	R2	**0.87**	0.81	0.77	0.62
Test	MSE	**1902.23**	2112.44	3012.34	3512.71
	RMSE	**43.61**	45.96	54.88	59.27
	MAE	**34.91**	36.07	44.66	48.78
	R2	**0.93**	0.92	0.89	0.87



(a) L5b-UP-D.

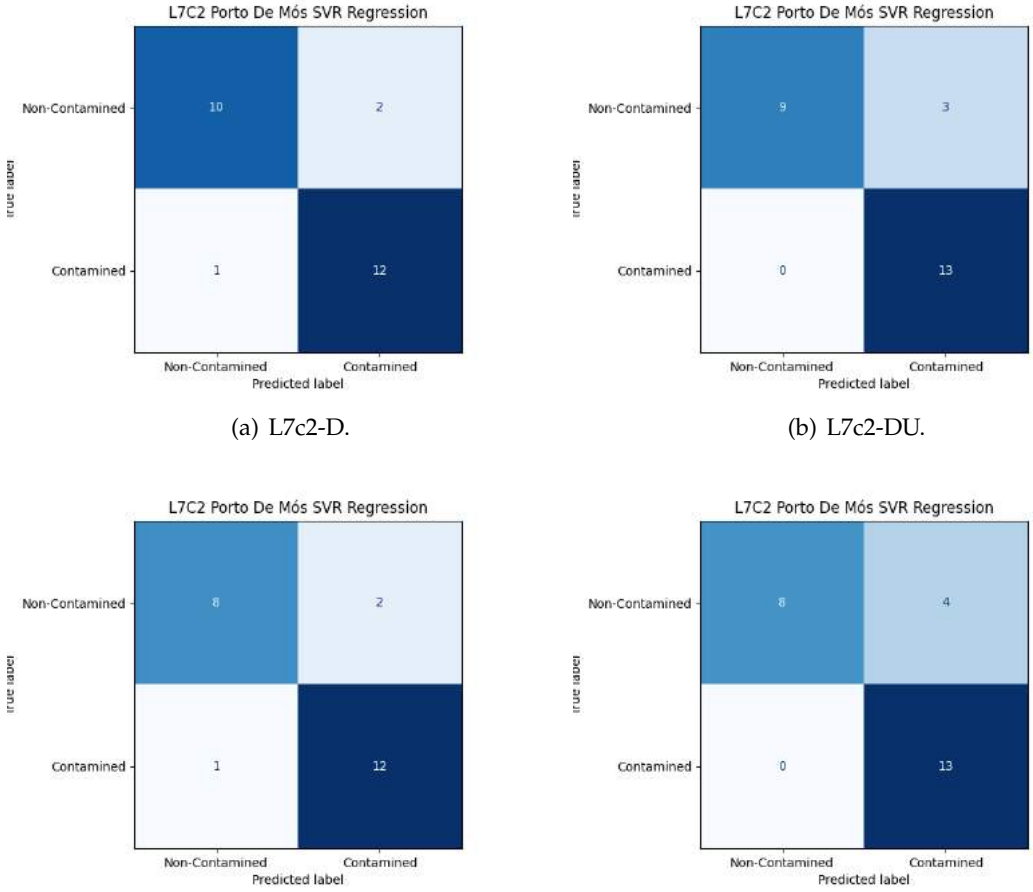
(b) L5b-UP-DU.

(c) L5b-DU-DM.

(d) L5b-UP-DMU.

Figure C.98: SVR L5b Caparica DSP predictions.

C.2.2.6 SVR Regression Upwelling L7c2 Porto de Mós



(a) L7c2-D.

(b) L7c2-DU.

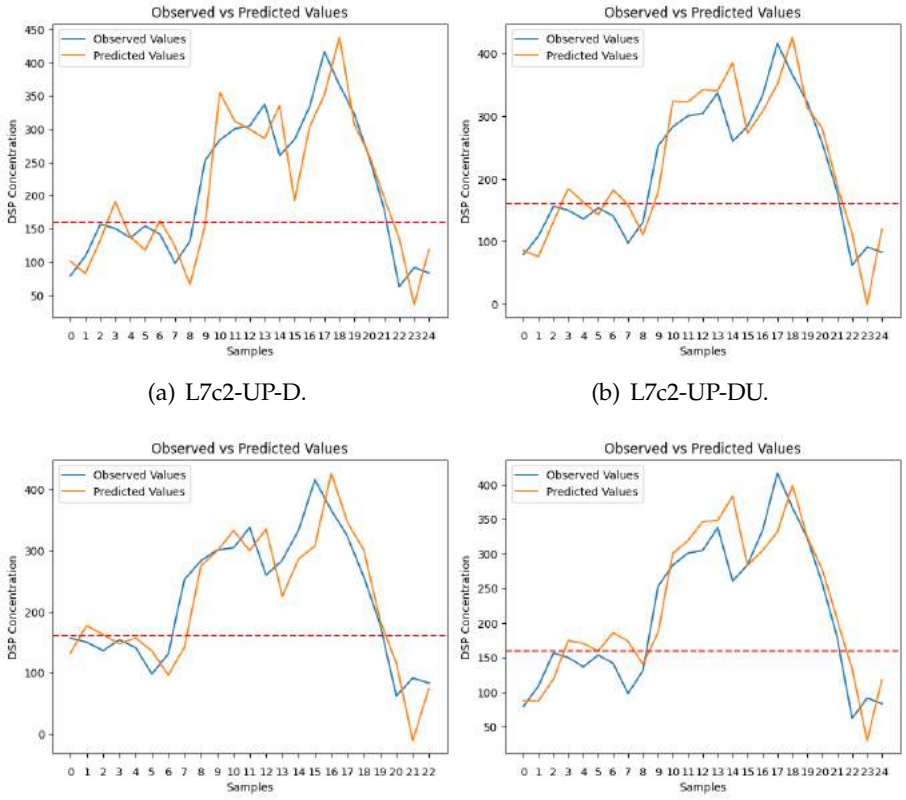
(c) L7c2-DM.

(d) L7c2-DMU.

Figure C.99: SVR Regression Confusion Matrices for L7c2 Porto de Mós.

Table C.24: SVR Upwelling Regression metrics for L7c2 Porto de Mós

Set	Metric	Dataset			
		D	DU	DM	DMU
Train	MSE	2603.64	1990.10	**1770.07**	2168.50
	RMSE	50.47	44.49	**41.87**	46.38
	MAE	25.06	25.02	**23.83**	28.04
	R2	0.80	0.85	**0.87**	0.83
Validation	MSE	**1279.28**	1952.11	2680.59	2705.20
	RMSE	**34.15**	41.57	49.14	50.45
	MAE	**24.97**	33.96	37.17	38.34
	R2	**0.92**	0.87	0.80	0.81
Test	MSE	2504.02	2255.84	2710.55	**2155.80**
	RMSE	50.04	47.50	52.06	**46.43**
	MAE	41.42	37.94	41.26	**36.01**
	R2	0.76	0.79	0.73	**0.80**



(a) L7c2-UP-D.

(b) L7c2-UP-DU.

(c) L7c2-DU-DM.

(d) L7c2-UP-DMU.

Figure C.100: SVR L7c2 Porto de Mós DSP predictions.

2023 Evaluating the role of environmental variables on shellfish biotoxin contamination via supe

Manuel Oliveira

NOVA SCHOOL OF SCIENCE & TECHNOLOGY