NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE
DEPARTMENT OF MATHEMATICS

JOÃO DANIEL LOPES DO CARMO SAMARÃO

BSc in Applied Mathematics to Technology and Business

UNVEILING THE GEARS IN SMALL-SCALE FISHERIES USING LANDINGS FROM A DIFFERENT COUNTRY: AN ML APPROACH

MASTER IN ANALYSIS AND ENGINEERING OF BIG DATA

NOVA University Lisbon
September, 2024

DEPARTMENT OF
COMPUTER SCIENCE
DEPARTMENT OF
MATHEMATICS

UNVEILING THE GEARS IN SMALL-SCALE FISHERIES USING LANDINGS FROM A DIFFERENT COUNTRY: AN ML APPROACH

JOÃO DANIEL LOPES DO CARMO SAMARÃO

BSc in Applied Mathematics to Technology and Business

Adviser: Marta Mega Rufino
Researcher, IPMA and CEAUL, Faculty of Sciences, University of Lisbon

Co-adviser: Paula Alexandra da Costa Amaral
Associate Professor, Nova School of Science and Technology, Nova University of Lisbon

Examination Committee

Chair: João Carlos Gomes Moura Pires
Associate Professor, Nova School of Science and Technology, Nova University of Lisbon

Rapporteur: Rui Alberto Pimenta Rodrígues
Assistant Professor, Nova School of Science and Technology, Nova University of Lisbon

MASTER IN ANALYSIS AND ENGINEERING OF BIG DATA

NOVA University Lisbon
September, 2024

Unveiling the gears in small-scale fisheries using landings from a different country: an ML approach

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.18) [33].

Acknowledgements

I would like to begin by thanking my supervisors, namely, Marta Rufino and Professor Paula Amaral, for their valuable feedback and insightful suggestions throughout the thesis, which enabled me to improve my work and gain a deeper understanding of the subject, leading to a more thorough and well-rounded analysis.

In addition, I am deeply thankful to both Instituto Português do Mar e da Atmosfera and Instituto Español de Oceongrafía for providing the data used in this thesis and for their willingness to clarify any questions regarding the dataset.

Lastly, I am especially grateful to my friends and family, including Deborah Andrade, Francisco Fonseca, Lisa Ye, Miguel Mota, Miguel Santos, and Raquel Almeida, for their invaluable feedback on the methodologies and concepts, helping ensure the explanations were clear to those without specific knowledge of the subjects. I am also deeply thankful for their constant encouragement and motivational support throughout this process.

”

“Do not go where the path may lead; go instead where there is no path and leave a trail.”

— Ralph Waldo Emerson
(Philosopher and Essayist)

Abstract

Small-scale fisheries (SSF) wield substantial influence on the maritime ecosystem, constituting 80% of the global fleet, 84% within the EU, and 87% of the Portuguese fleet. While these fisheries exert such an important role, only recently have been established proposals to improve the management of SSF, which contrary to Large-Scale fisheries, operate without particular obligations.

Within the proposals, logbooks containing official records register information about the species caught and its respective weight in kilograms, and fishing gear operated, however, compliance with logbook requirements remains subject to the fishermen's intention, which makes it challenging to create a comprehensive analysis of these records.

Moreover, the Portuguese fleet often operates with different gears, but these are not validated when used during a boat trip. The only information available is which gear licences were provided, making it complicated to understand how these gears are related to the species caught and their behaviour in the Portuguese fishing grounds.

Nonetheless, in the current work, a dataset with the gears validated was provided by the Instituto Español de Oceanografía (IEO), which was used to implement supervised machine learning models. These models performed excellently on unseen data, whereas XGBoost (XGB) was the highlight by having a better recall than Random Forest (RF)(97% - XGB and 96% - RF), using the species caught weight in kilograms, the length of the boat, the taxa richness, the month of the landing, and the total species caught weight in kilograms as variables. XGB was then applied to the Portuguese dataset which validation was not provided but had an overall coherence with the main licenses of 81% and a coherence with fisheries experts of 82%.

After testing different models, Shapley Additive Explanations and Logistic Regression were used to better understand the relationship between gear classification and the selected variables. This analysis provided insightful results. Dredges were associated with catching Chamlea, Donax, and Bolinus. Pots & Traps were linked to Octopus, Sepia, and more activity early in the year. Gillnets were identified by catching Merluccius, Diplodus, and showing higher species richness. Trammel Nets stood out for higher species richness, catching Mullus, Sepia, Pegusa, and being used by smaller boats. Hand & Pole Lines were

associated with catching Mullus, Pagrus, Octopus, and Raja, and lower species richness. Longlines were characterized by catching Pagrus, Pagellus, Phycis, and Muraena. Otter Bottom Trawlers were linked to larger boats, higher species richness, and catching Lophius, Octopus, Mullus, Trachurus, Sepia, Microchirus, Loligo, Solea, and Merluccius. Finally, Purse Seiners were characterized by a high total catch weight and species like Scomber, Trachurus, Euthynnus, Mullus, Pomadasys, and Sardina.

Keywords: Small-Scale Fisheries, Machine Learning, Landing Profiles, Shapley Additive Explanations, Fisheries Management

Resumo

A Pequena Pesca ou a Perca Artesanal exerce uma influência substancial no ecossistema marinho, constituindo 80% da frota global, 84% na União Europeia e 87% da frota Portuguesa. Embora estas pescarias desempenhem um papel importante, somente recentemente foram estabelecidas propostas para melhorar a gestão das mesmas, que, ao contrário das pescarias de grande escala, operam sem obrigações específicas.

Dentro das propostas, os *logbooks*, onde são registadas informações oficiais, como espécies capturadas e o seu respetivo peso em quilogramas, e artes de pesca utilizadas, ainda dependem das intenções dos pescadores, o que torna desafiante a implementação de uma análise mais abrangente destes registos.

Além disso, a frota portuguesa muitas vezes opera com diferentes artes de pesca, sendo que apenas as licenças são fornecidas (ou seja, as artes não são validadas nos portos), o que torna complicado entender como estas estão relacionadas com as espécies capturadas e seus comportamentos nas áreas de pesca portuguesas.

No entanto, no presente trabalho, foi fornecido um conjunto de dados com as artes de pesca validadas pelo Instituto Español de Oceanografía (IEO), que foi utilizado para implementar modelos de aprendizagem automática supervisionada. Estes modelos apresentaram um desempenho excelente em dados não vistos, destacando-se o XGBoost (XGB), que obteve um melhor recall do que o Random Forest (RF) (97% - XGB e 96% - RF), utilizando como variáveis o peso das espécies capturadas em quilogramas, o comprimento da embarcação, a diversidade de espécies, o mês do desembarque, e o peso total das espécies capturadas em quilogramas. O XGB foi então aplicado ao conjunto de dados português, cuja validação não foi fornecida, mas apresentou uma coerência global com as principais licenças de 81% e uma coerência com os especialistas em pesca de 82%.

Após testar diferentes modelos, as Explicações Aditivas de Shapley e a Regressão Logística foram usadas para entender melhor a relação entre a classificação do tipo de artes de pesca e as variáveis selecionadas. Esta análise trouxe resultados interessantes. Os arrastos mecânicos foram associados à captura de Chamelea, Donax e Bolinus. Armadilhas e Covos foram relacionados com a captura de Polvo, Sépia, e com mais atividade no início do ano. As redes de emalhar foram identificadas pela captura de Merluccius, Diplodus,

e por apresentarem maior riqueza de espécies. As redes tresmalho destacaram-se pela maior riqueza de espécies, captura de Mullus, Sépia, Pegusa, e por serem operadas por embarcações mais pequenas. As linhas de mão e cana foram associadas à captura de Mullus, Pagrus, Polvo e Raja, e menor riqueza de espécies. As palangres foram caracterizadas pela captura de Pagrus, Pagellus, Phycis e Muraena. As redes de arrasto de fundo estavam ligadas a barcos maiores, maior riqueza de espécies, e captura de Lophius, Polvo, Mullus, Trachurus, Sépia, Microchirus, Loligo, Solea e Merluccius. Finalmente, os cercadores caracterizaram-se por um grande peso total de captura e pela captura de espécies como Scomber, Trachurus, Euthynnus, Mullus, Pomadasys e Sardina.

Palavras-chave: Pequena pesca, Aprendizagem Automática, Perfis de desembarque, Shapley Additive Explanations, Gestão de Pescarias

Contents

List of Figures

List of Tables

1

Introduction

Small-scale fisheries (SSF) wield substantial influence on the maritime ecosystem, constituting 80% of the global fleet, 84% within the EU, and 87% of the Portuguese fleet [23, 49, 53]. Differing from Large-scale fisheries (LSF), SSF vessels, with an overall length (LOA) of less than 12 meters, have operated without specific obligations such as fishing authorisations, landing declarations, and sales notes [12].

Recent proposals aim to amend Council Regulation (EC) No 1224/2009 (2023) to improve management practices for SSF. These measures include mandatory tracking systems for all vessels and electronic reporting of logbooks (an official record of events during the voyage of a ship, e.g., species caught, fished area, and gear used), including those with LOA < 12m. The implementation of these measures would significantly enhance the efficacy in the management of SSF.

Within the proposed regulation, compliance with logbook requirements remains subject to the fishermen's credibility, making it challenging to create a comprehensive analysis of these records. Nevertheless, the insights derived from these records prove valuable for implementing bio-economic models, facilitating the analysis of catches and their economic determinants. Additionally, it contributes to overseeing stock depletion, examining fishing gears and relating it to catch composition.

Logbook data can also contribute to additional research through the analysis of the métier, which should reflect the fishing intention such as the species target, the geographic area visited, and the gear used or, as stated in (EU) 2016/1251, métier is a group of fishing operations targeting a similar assemblage of species, using similar gear, during the same period of the year and/or within the same area and which are characterised by a similar exploitation pattern [36].

By combining all these factors (métier and logbook data) it is possible to assess their impact on profitability and long-term sustainability and measures such as the landing per unit of effort (LPUE) and the catch per unit effort (CPUE) can be associated with each métier [32, 57]. Nonetheless, to study the connection between the métier and these coefficients it is essential to have information about the fishing gear and its relationship with the species.

Therefore, this thesis aims to analyse the landing data (fishing vessel offloads data, e.g., species caught) of the Portuguese Small-Scale Fisheries by unveiling the gears operated, understanding which species are targeted by each gear, and identifying additional patterns, such as seasonal variations. However, in addition to what was previously mentioned about logbook data not being reliable enough, Portuguese Fisheries provide fishing licenses that are not validated within the trips which is an issue as these fisheries often operate more than one gear [65].

Thus, to address this challenge, a subset of landing data from vessels in Algarve (south of Portugal) was retrieved, and in collaboration with the Instituto Español de Oceanografía (IEO) a validated subset of vessels from South of Spain with similar characteristics within the fleets and fishing grounds was provided to implement and train supervised algorithms that would be used to classify the Portuguese landings and further understand the relationship between the gears, species, and additional features. Lastly, the results obtained from the classification of the Portuguese landings will be validated by fisheries experts to obtain a metric of reliability.

Thereafter, to achieve the established objective, the thesis is organised as follows: A chapter that goes through previous works related to this subject and explores what could be used or not in the current work (Chapter 2). A data section where it is explained how the datasets from IPMA and IEO were unified, followed by an exploratory analysis and how the data was organised to be used to further implement models (Chapter 3). A chapter that introduces and explains the methodologies that will be used (Chapter 4). Afterwards, the results will be displayed in Chapter 5. A discussion about the thesis and future work is presented in Chapter 6 and finally, a conclusion about the thesis is provided in Chapter 7.

2

State of the Art

Few studies cover the relationship between gear or métier and species on Small-Scale Fisheries (SSF) through logbook/landing data due to its reliability as described in the previous chapter (1) and of those few, Palmer et al. (2017) studied Mallorca's (Western Mediterranean) small-scale fleet demonstrating how fishers' expertise can be combined with electronic registering landings to ascribe the métiers practised within the fleet [45]. The authors implemented several machine learning algorithms, whereas Ibk (an algorithm that implements a k-nearest neighbour classifier) demonstrated to achieve greater performances (an accuracy of 99.90% over the cross-validation procedure). The authors opted for a different strategy, instead of categorizing all landing reports to determine their respective métier, it was applied a binary classification for each métier individually, whereas the outcome was whether or not a given report belongs to a specific métier. Furthermore, despite presenting good results on the cross-validation procedure, the authors decided to apply the models to all the data without relying on a testing set.

However, as the objective is to have a framework with a single algorithm that allows us to identify different gears given any landing record, this approach was not considered in the current thesis.

Furthermore, some authors used both SSF and Large Scale Fisheries (LSF) data to study the fishing gears, where Szynaka et al. (2021) analysed the Portuguese landings in Algarve with fleet varying from below 10 meters to above 16 meters (i.e., including both SSF and LSF) [65]. This analysis was conducted using logbook and VMS data. The authors performed a Multivariate Regression Tree to evaluate different factors on the species caught including gear and seasonal variations, highlighting that the gear was a crucial variable to distinguish different species. Despite using a curious and innovative approach to define which species are targeted by each gear, in the Portuguese Landings, we cannot have the gear licenses as predictive variables as it cannot be confirmed if the gear in the license was used. Therefore, as promising as this approach could be to take insights into the relationship between species and gears, in the current thesis it is not possible to be applied.

Moreover, Russo et al (2011) uses Vessel Monitoring Systems (VMS) to identify fifteen

possible métiers [55]. The authors preferred to use tracking data as logbook data is usually incomplete or biased. Nonetheless, the authors implemented a multilayer perceptrons network with thirty-three variables, considering the gears, classes of speed, depth and heading (angle through which the boat was bearing). This approach resulted in a 94% accuracy on the test dataset, which makes VMS data reliable for providing information on vessel activity. Despite excellent performance, in the current thesis, we cannot implement a similar approach as there is no VMS data available, however, as multilayer perceptrons are a methodology easier to implement, these were considered in the first tests.

Additioanlly, Russo et al. (2016) combined logbook and Vessel Monitoring System (VMS) data to characterize the composition of landings per métier for the Italian fleet (only Large Scale Fisheries data was used, LOA > 12m) by extracting and predicting relevant patterns [54]. To achieve this the authors, first identified the fishing gears used by applying Self-Organized Maps (SOM), which are a particular type of unsupervised Artificial Neural Network (ANN) that produces a synthetic representation of the information in the input data by approximating their probability density function, reaching an accuracy of 86% over nine gears, whereas the most difficult to identify were Trammel Nets (GTR), Purse Seines (PS) and Pelagic Pair Trawlers (PTM) (Description of the gears can be found in [13]). The approach used by the authors seems very promising, however, SSF do not have VMS, making it impossible to match the landing and tracking data, and therefore, replicate what was done in the authors' work. By not having geo-spatial information available and relying only on the landing data it was decided to not implement this procedure in the current thesis as there is no guarantee of achieving similar results.

Aside from applying tracking data to identify métiers, some authors used this data to identify fishing gears as the current objective of this thesis. Despite not being able to apply these approaches it is important to highlight some of the work that have been already developed at the fishing gear identification level.

Rodriguez-Albala, et al. (2024) used GPS tracking data to identify seven fishing gears. Analyzing this data as a time sequence the authors applied common supervised learning algorithms achieving accuracies up to 90% [51]. Other authors applied Convolutional Neural Networks (CNNs) using automatic identification system-based (AIS) trajectory data of fishing ships [31]. The authors' work aimed to identify six fishing gears and achieved a total performance of the day-wise performance index of 0.963. this metric was not applied in the current thesis as it is related to time series analysis, yet the authors state that it is comparable to the accuracy of the models.

Moreover, Marzuki et al. (2015, 2017) implemented supervised and unsupervised machine-learning techniques to identify fishing gear through VMS data. Using the supervised learning approaches, namely, Random Forest and Support Vector Machines, the authors achieved a correct recognition rate of 94.59%, while using unsupervised techniques the authors achieved a recognition rate of 97% [37, 38].

Carlos et al. (2021) used the supervised autoencoder dimensional reduction algorithm to identify more than five fishing gears. This algorithm had better performances than any

other algorithm used in the author's work achieving an accuracy of 95% [9].

Despite numerous studies achieving excellent performance in fishing gear classification, one key advantage of analyzing landing data is identifying which species are most associated with each type of fishing gear. However, few studies have explored gear classification while emphasising the species strongly linked to particular fishing methods. Most existing research tends to focus on specific gears and species, investigating aspects such as bycatch analysis (unwanted fish and other marine creatures trapped by commercial fishing nets during fishing for a different species), net length, or forecasting the quantity of fish caught in terms of kilograms. Numerous studies have examined these topics, including works such as [2, 4, 6, 16, 19, 47, 50, 60]. While many more could be referenced, these are provided for brevity, with additional studies to be cited in the discussion section.

Nonetheless, to study the relationship between fishing gears, species, and other variables, a tool is needed to interpret the decision-making process of the models. Models like Random Forest and XGBoost typically provide feature importance, but this is based on how often a feature is used to split a node in the decision trees. In contrast, more interpretable models like Logistic Regression or K-Nearest Neighbors offer coefficients for each feature, making it easier to understand how each variable contributes to the final decision. Moreover, with advancements in new technologies, more complex models are being developed that offer improved accuracy but often come at the cost of reduced interpretability.

Considering this, several methodologies such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), have been developed to provide insights into these models' decisions, improving the interpretability of these models.

SHAP provides consistent and theoretically sound explanations by considering all possible feature interactions, providing global and local interpretability, and making it versatile for various model types. The algorithm has been applied in several fields of science, such as gold prices forecast and freight truck-related crashes [30, 67].

LIME, on the other hand, offers a more local perspective, generating explanations for individual predictions by approximating the model locally, which is ideal for understanding individual predictions in simpler models. This methodology has been applied in fields such as medicine and pharmaceutics [17, 27, 69].

In the current thesis, SHAP was used to interpret the black-box models implemented as the objective is to understand which features are related to the fishing gears globally and individually, contrary to LIME which only offers an individual analysis. Logistic Regression was also considered for a straightforward interpretation of the model coefficients.

In summary, significant progress has been made in fishing gear classification and related analyses, with numerous studies offering valuable insights. However, key areas, such as the integration of species-gear relationships and model interpretability, remain

insufficiently explored. As technology evolves, the challenge of balancing model complexity with interpretability becomes increasingly relevant. These gaps present opportunities for further research, which this study aims to address by exploring both the classification of fishing gear and the species and other variables most closely associated with each gear type.

3

Data

The data consists of official landings from the south of Spain (provided by Instituto Español de Oceanografía, IEO) and from the south of Portugal (provided by the Portuguese Instituto of the Sea and Atmosphere, IPMA, but obtained by Direção-Geral de Recursos Naturais, Segurança e Serviços Marítimos, DGRM).

The data provided by IEO includes 808 vessels from trips between January 2021 and January 2022 (52063 trips). Each trip can have more than one landed profile entry as more than one species can be caught within the trip, registering 211 distinct species (by scientific name) caught over these dates. These species were captured by 9 gears and information about kg and the commercial values of each fishing trip were reported.

The Portuguese landings data shows 753 vessels (overall length <= 15m) using 13 gear licenses (including main and subsidiary licenses, in Algarve) in 2021. These vessels reported 60396 trips over the year with information on the landed species, along with the weight (in kg) and the respective commercial value. Similarly to the data from IEO, a boat trip can have more than one landed profile entry capturing more than one species in a single trip, registering 409137 entries. From these entries, 177 distinct species were reported (by scientific name). Note that more species may be caught but not reported or even returned to the ocean mid-trip as the species might not hold commercial value (bycatch).

Besides the species-related information above, both datasets include features such as the date of the landing, port, the gear (only license by vessel in the IPMA dataset, validated gear - IEO dataset), and Season (Winter, Spring, Summer, and Fall). Boat characteristics for each landed profile such as length of the boat (LOA), length between perpendiculars (LBP), and power of the main engine were obtained from the EU boat registry.

3.1 Establishing a Unified Framework

One of the objectives of this thesis is to use the validated gears of the IEO's dataset to implement an algorithm that allows us to classify the Portuguese gears in Algarve. Thus, a first step was conducted to match the gears that are present in both datasets, which

are Dredges (DRB), Pots and Traps (FPO), Gillnets (GNS), Trammel Nets (GTR), Hand and Pole Lines (LHP), Longlines (LLS), Bottom otter twallers (OTB), and Purse seines (PS)(more detail about the gear in [13] and respective illustrations can be found in Figure 3.1).

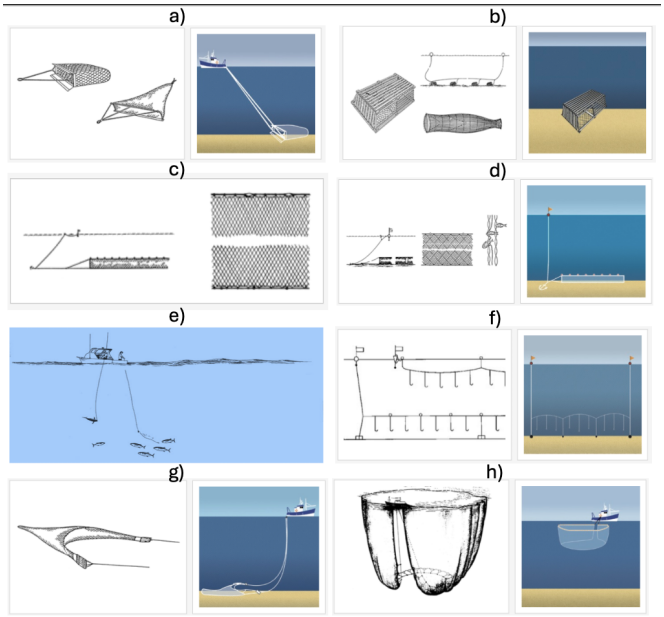


Figure 3.1: Fishing gear illustrations: a) DRB; b) FPO; c) GNS; d) GTR; e) LHP; f) LLS; g) OTB; h) PS

Due to gears that were not used by the Spanish fleet (or at least validated) and used by the Portuguese fleet, and vice-versa, the match between the gear made a decrease in Portuguese vessels by 2% and by 2.6% on the Spanish vessels and in the number of trips reported by 1.59% and 4%, respectively. Relatively to the taxa, only four were removed by this match on the IEO's dataset.

Furthermore, to establish a unified dataset between the two institutes the taxa of both countries were connected to the WoRMS database to obtain the most recent nomenclature and retrieve species' Aphia ID, Family, Genus and Order [1]. Additionally, to increase the match between datasets, species were reclassified into a broader taxa level (or group), in all cases to the respective genera (Genus). By changing all species to the respective genera, a new total number of species was defined, 133 genera within the Portuguese data and 144 within the Spanish. Before matching genera between datasets, species that were caught in less than 10 trips and trips with a total catch <0.5kg were removed to avoid noisy landing reports in further analysis (similar steps were conducted in [32, 54]), these restrictions decreased the number of trips of the Portuguese fleet to 59434 (<1%), and the number of species genera to 108 and 115, for IPMA and IEO, respectively. Additionally, as the objective is to study the gears within the Small Scale Fisheries (SSF) fleet, the boat trips were filtered to boats with lengths of less or equal to 12 meters. This reduced the number of boats to 638 (6.1%) and 710 (9.8%), which corresponds to a decrease in the number of

trips 7.4% (55041 trips) and 12.5% (45566 trips), respectively for IPMA and IEO. As for the species, 1 was removed when applying the filter within the IPMA dataset, while 4 were removed from the IEO dataset.

Lastly, the match between the species genera was performed, whereas Table 3.1 illustrates the final number of vessels, trips, entries and species.

Table 3.1: Summary of the number of vessels, trips, entries, and species after establishing a unified dataset

	IEO	IPMA
#Vessels	655	693
#Trips	38929	54315
#Entries	107329	337902
#Genera	71	71

3.2 Data Processing

After cleaning the data and selecting the variables that should be used in the following steps, it is important to organise and process the data that will be given to the models. At the current state, the data is arranged by entry (remembering that in a single boat trip, it can be registered several catches, i.e., entries), however, to study and perform further analysis it is desired to have the data organised by boat trip. Therefore, after defining a boat trip by boat ID and date of the landing, Figure 3.2 illustrates how the data was transformed from information by entry to information by boat trip.

Trip	Date	Species	Kg	...	Season	LOA	Gear
Trip_1	13/02/2020	Species_1	10.7	...	Winter	7.10	DRB
Trip_1	13/02/2020	Species_2	5.7	...	Winter	7.10	DRB
Trip_2	19/02/2020	Species_1	1.5	...	Winter	11	PS
Trip_1	13/02/2020	Species_2	2.4	...	Winter	7.10	DRB
...	...	...	...	...	...	...	...

Trip	Species_1/kg	Species_2/kg	...	Species_n/kg	Season	LOA	Gear
Trip_1	10.7	8.1	...	0	Winter	7.10	DRB
Trip_2	1.7	0	...	0	Winter	11	PS
...	...	...	...	...	...	...	...
Trip_n	...	...	...	...	...	...	...

Figure 3.2: Landing report entries transformed as information by boat trip

3.3 Exploratory Data Analysis

After cleaning and processing the data it is essential to understand more about our variables and how they can contribute to the established objectives. Furthermore, it is important to comprehend how similar or different the fleets and season behaviours are between the two datasets.

Figure 3.3 shows that half of the Spanish fleet is larger than 8.25 meters and that the distribution is almost even as the median is slightly above the middle of the box. Looking at the histogram, there are more trips using boats with a length between 6 and 9 meters.

Differently, more than half of the Portuguese fleet (close to 75%) is greater than 7 meters as the median line is close to the first quantile. The histogram illustrates this pattern as there are more trips using boats with a length above 6 meters. Nonetheless, both fleets present similar characteristics.

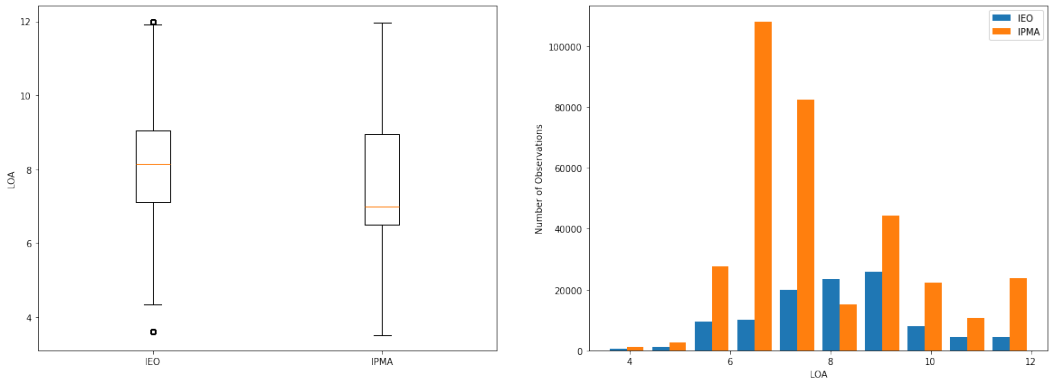


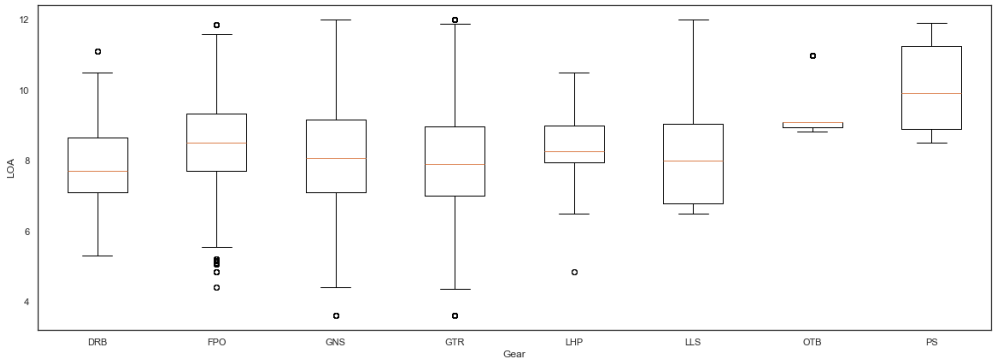
Figure 3.3: Spanish and Portuguese fleets overall length (LOA)

The following analysis is only done for the IEO dataset as we are analysing the relationship with the gears and, as mentioned previously, within the Portuguese fleet is not possible to guarantee that the gears used were the ones provided in the licenses.

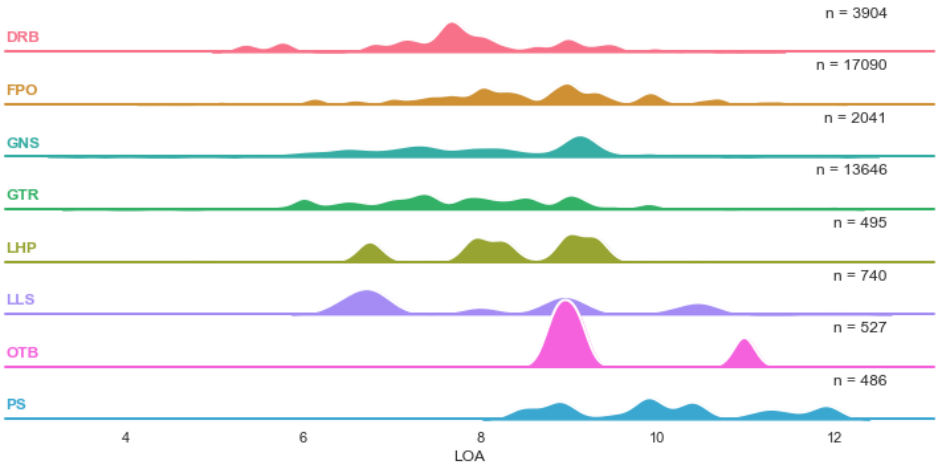
Figure 3.4 shows the overall length (LOA) of the boats by the different fisheries. It is possible to highlight that fisheries that are fishing with OTB and PS usually operate with larger boats. LLS fisheries tend to operate with boats between 6 meters and 12 meters, having more vessels with a LOA below 8 meters. Moreover, FPO, GNS, GTR, and LHP all operate with boats with a length between 6 and 10 meters. The smallest fleets are from DRB which do to present any boat larger than 11 meters excluding the outliers. In contrast, this gear is the only one used by boats with lengths below 6 meters, excluding the outliers.

Another important aspect to consider is the season variability, which has been demonstrated in previous works to greatly impact which species are caught, and therefore, impacts the usage of some gears [32, 65]. Figure 3.5 illustrates how the kilograms caught by the number of trips vary with the months and the gear. Additionally, as this measure will change depending on the gear, these values were normalized at the gear level (by subtracting the minimum value observed and dividing it by the range of values). Therefore, values closer to zero characterise months with lower catch rates (in kg) by the number of trips of a specific gear. On the other hand, values closer to one represent months with higher catches. Months where no activity was registered are blank (white).

The beginning of the year illustrates more activity (January to March) where DRB, FPO, LLS, and PS have higher catches by boat trip. On the other hand, GNS, GTR, LHP, and OTB present lower catches. In the following months (April to July), the fisheries register an increase or maintain the catch rate except for DRB, which is not active during May and June but reaches its highest catch rate in July. Between May and July, all fisheries register their maximum catch rate, excluding LLS. After August, there is a decrease in the



(a)



(b)

Figure 3.4: Length of the boats per each gear - (a) Boxplot (b) Density plots

catches by boat trip in every fishery except for PS, which has high catch rates over the year besides November and December. As for the other fisheries, few increase their catches again at the end of the year, namely, DRB, GNS, GTR, and LLS. Overall, the first half of the year has more fishing activity, whereas PS is the only fishery that maintains its fishing effort equally throughout the year except for November and December.

The catches by month illustrated a few patterns where more fishing activity occurs, demonstrating to be a variable to consider when implementing the machine learning models. However, if we look at this information from another perspective, it is possible to gain more and different insights. Figure 3.6 shows how the kilograms caught by the number of trips vary with the seasons and the gear (Fall: September, October, and November; Spring: March, April, and May; Summer: June, July, and August; Winter: December, January, and February).

Spring is the season where more fishing activity is registered, as the maximum accumulative catches of most fisheries occur in this season. On the other hand, summer illustrates fewer catches despite some fisheries presenting the highest catch rates in June

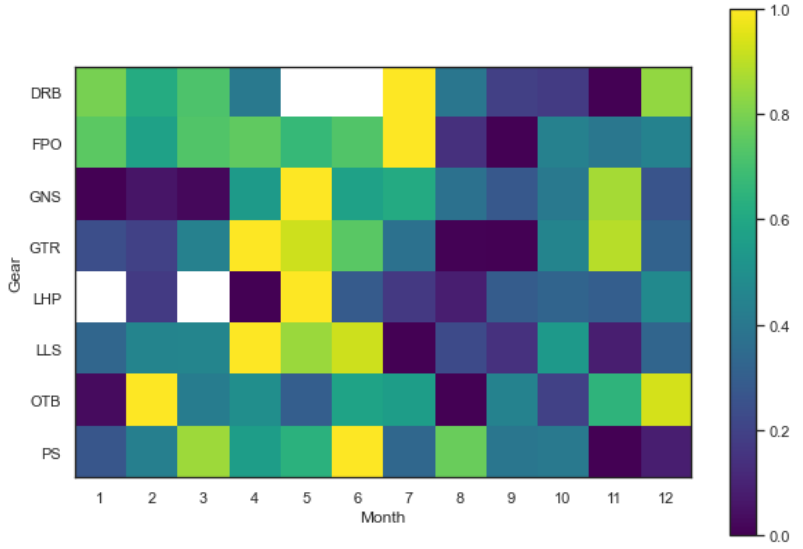


Figure 3.5: Monthly kilograms caught by number of trips

and July (DRB, FPO, and PS). At the end of the year, where monthly represents lower catches, when aggregating these catches by season, fall is one of the seasons with more activity. Winter equally demonstrates higher activity but is more present at the beginning of the year, as illustrated in the previous Figure 3.5.

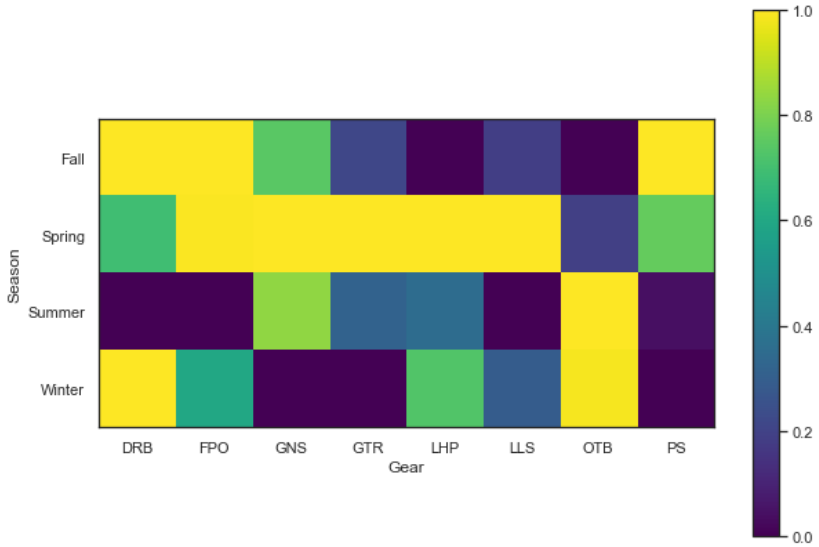


Figure 3.6: Seasonal kilogram caught by number of trips

Moreover, one of the variables that could help distinguish the different fisheries is the number of species caught within the trip (i.e., the taxa richness - S), as some fisheries may capture single species groups (broader taxonomic levels such as genera). Thus, Table 3.2 illustrates the mean and mode of the taxa richness of each fishery, where it is possible to analyse that on average DRB, FPO, and LHP mostly caught one species group and OTB,

for example, approximately 8 species groups per trip. Nonetheless, looking at the mode, most of the fisheries tend to catch one species per trip except GNS, GTR, and OTB.

Table 3.2: Mean and Mode Taxa Richness (S) per gear

Gear	DRB	FPO	GNS	GTR	LHP	LLS	OTB	PS
Mean S	1.03	1.00	4.27	4.80	1.30	3.97	7.84	1.97
Mode S	1	1	3	3	1	1	9	1

4

Methods

As mentioned in previous sections, supervised machine learning was used to implement a methodology that would allow us to identify the Spanish fishing gear and further apply it to the Portuguese fleet in Algarve. Moreover, to understand the relationship between the variables and the respective gears Shapley additive explanations and Logistic Regression were applied.

4.1 Supervised Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence which can be organised into three categories: (1) Supervised Learning, (2) Unsupervised Learning, and (3) Semi-Supervised Learning [26]. Of these, the current thesis focuses on the first one - Supervised Learning.

Supervised Machine Learning can be applied when there is a dataset where the data has a target (a response variable) that has been validated (ground truth). In each observation (each row of the dataset) there will be several features (each column of the dataset) related to it, whereas in this case, one of them is the target, which is the variable that we want to study and predict. An example could be derived from Figure 3.2 where each observation is a boat trip and it has features such as the kg caught of each species, season, LOA, and lastly the gear (target variable). Therefore, the objective of supervised machine learning is to use all the features of each observation to find a pattern that will describe the target variable, and generate a prediction.

Within supervised ML the predictions can be quantitative (continuous numerical value) or categorical, depending on which one it is, it is possible to perform two types of tasks: Regression and Classification. Regression is the task of predicting continuous numerical values, whereas the output of the model is a continuous numerical value too. As for classification, it will be used when the target variable is categorical, whereas, unlike regression, the outcome of the model will be a vector of probabilities of an observation belonging to each category (or class). The response of the model will be the class with the highest probability or odds.

Usually, to implement these models we have training and testing sets (samplings of the main dataset). The supervised model learns from the training data and is optimised on it (usually using a cross-validation procedure, details on section 4.2). Further, its final performance is evaluated on the unseen data (the testing set). Based on the results, it is decided if the model should be deployed in practice, or if it would be adequate to classify the Portuguese landings, in this thesis context.

The models implemented in this work will focus on classification tasks, aiming to identify the specific fishing gear used during each boat trip. Therefore, the following sections will provide examples exclusively related to classification scenarios.

4.1.1 Random Forest

Random Forest is an algorithm that creates a "forest“ of many decision trees. Thus, to understand the concept behind this algorithm, it is required to clarify what a Decision Tree is.

Decision Tree

To illustrate the underlying idea of a decision tree, we can look at Figure 4.1, which shows what a simple tree with a maximum depth of 2, for classifying the gear used within a boat trip could be considering that the classes are in the following order: DRB, FPO, GNS, GTR, LHP, LLS, and OTB (looking at the *values* attribute at the first node, from left to right, of the nodes at depth = 2).

If we were to classify a trip of the landing profiles, looking at the top of the figure (1^{st} node, depth = 0), and this trip would have less than or equal to 9.165 kg of octopus it would go to the left of the tree, otherwise to the right and so on through the other layers of the tree until reaching the max depth (the last layer). The last layer is where the decision is made, considering that after the previous decision, we went to the left again, i.e., the kg of Donax would be less or equal to 0.67, and the gear decision would end up as GTR.

Therefore, a decision tree is like a flowchart that asks a series of questions about the features of the trip and eventually assigns it to one of the categories (DRB, FPO, or another gear). This procedure does not require a lot of data preparation, in particular, it does not need feature scaling or centring at all, which illustrates one of the unique qualities of this algorithm [26].

Furthermore, each node *samples* attribute (which is only illustrated for the node of the example above) counts how many training instances it applies to. For example, 16454 entries have less than 9.165 kg of octopus caught (depth = 1, the left node) and no more than 0.67 kg of Donax (depth = 2, 1^{st} node from left to right). As for the *value* attribute, it illustrates the number of instances of each class in that node. In this node, we would have 965, 197, 1636, 10697, 58, 646, 941, and 1314 samples of each class respectively.

Nonetheless, we need to look at these last nodes to estimate the probability of an observation belonging to each class. Let's consider the last example, the likelihood of a

trip using each gear would be $\frac{965}{16454}$, $\frac{497}{16454}$, $\frac{1636}{16454}$, $\frac{10697}{16454}$, $\frac{58}{16454}$, $\frac{646}{16454}$, $\frac{941}{16454}$, and $\frac{1314}{16454}$, which means the gear that was probably used is GTR, with a probability of 0.65.

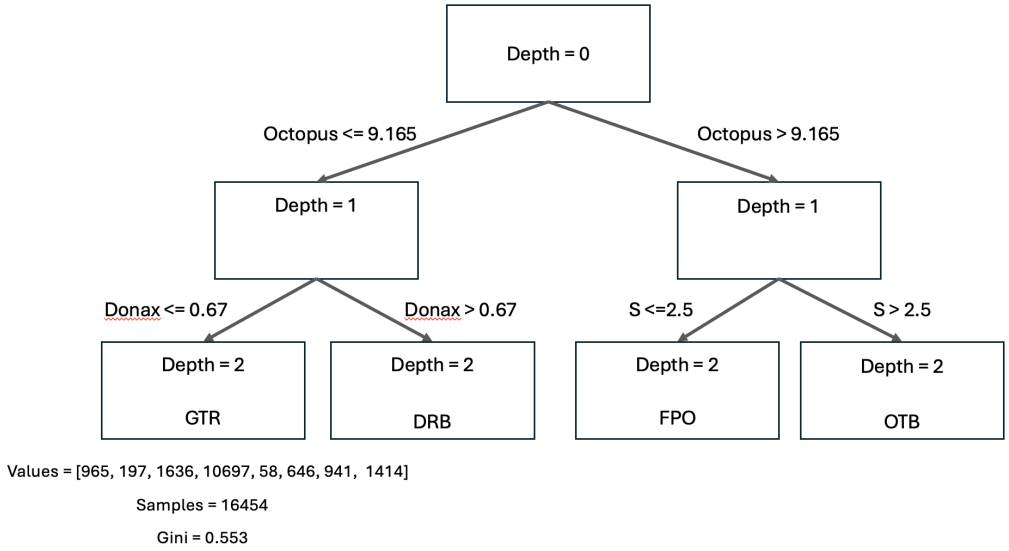


Figure 4.1: Decision Tree example with data from landing profiles

Lastly, the *Gini* attribute calculates the impurity of the node. If the measure is equal to zero, a node is pure, i.e., if all training instances belong to the same class within the node. Contrarily, if it is close to one, the instances with a specific characteristic could belong to several classes. The Gini Impurity G_i of the i-th node is estimated following Equation 4.1

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \tag{4.1}$$

where $p_{i,k}$ is the ratio of class k instances among the training instances in the i-th node.

There are other impurity measures such as Entropy (Equation 4.2). Usually, the Gini impurity is used, but the entropy impurity could be applied instead. The concept of entropy was introduced in thermodynamics as a measure of molecular disorder: entropy approaches zero when molecules are still and well-ordered. Later this concept was spread to a wide variety of domains. In Machine Learning, it is frequently used as an impurity measure, the entropy is zero when it contains instances of only one class.

$$H_i = -\sum_{k=1}^{n} p_{i,k}^2 log_2(p_{i,k}), p_{i,k} \neq 0 \tag{4.2}$$

where $p_{i,k}$ is the ratio of class k instances among the training instances in the i-th node.

Nevertheless, using Gini or Entropy does not make a big difference as it usually leads to similar trees. Gini impurity is slightly faster to compute, so it is a good default, and

therefore, used in the current work. However, when they differ, Gini impurity tends to isolate the most frequent class in its own branch of the tree, while entropy tends to produce slightly more balanced trees [26].

Randomness and Decision Making

The "random" in Random Forest comes from the fact that each decision tree is slightly different. The algorithm randomly selects a subset of the input data and a subset of the features for each tree. Furthermore, the selection of the input data for each tree can be performed with replacement, i.e., re-use the samples used in other nodes (this approach is called Bagging) or performed without replacement, i.e., not using the same samples used in other nodes (this approach is called Pasting).

Once every tree is built and trained, it is possible to make a final prediction by selecting the most frequent class among all the trees. As the final model results from an assembly of all trees produced it creates a diverse set of decision trees that can capture different patterns and reduce overfitting ([26]).

Feature Importance

Another aspect of Random Forests is the measurement of the relative importance of each feature. The library used in the current work to implement the Random Forest does this calculation by analysing how much the tree nodes that use a specific feature reduce the impurity on average across all trees in the forest, i.e, it is a weighted average, where each node's weight is equal to the number of training samples that are associated with it.

4.1.2 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm that combines multiple weak predictive models, usually in the form of decision trees (others could be used such as linear models, e.g., Logistic Regression). It follows an additive training process, where each subsequent model focuses on reducing the errors made by the previous models, similar to Gradient Boosting (see [25]). The key idea behind XGBoost is to gradually enhance the overall prediction accuracy by training weak models sequentially. During this process, XGBoost uses gradient descent to optimize a specified loss function (like log-loss for classification, Equation 4.3). Each new model is trained to minimize the residual errors or the differences between the predicted and actual values from the previous model. This iterative correction helps XGBoost improve its predictions over time [10, 26].

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N} y_i ln(p_i) + (1 - y_i) ln(1 - p_i) \tag{4.3}$$

where N is the total number of observations, y_i is the actual/true value for the i-th observation and p_i is the prediction probability for the i-th observation.

Initially, all examples are given equal weights. However, as the models are trained, these weights are adjusted to emphasize examples that are more challenging to classify or predict correctly. For instance, if a model underestimates the classification of trips operating Gillnets, the next model will focus more on reducing the error for these particular trips. By doing so, XGBoost can effectively learn from the mistakes made in previous models and refine its predictions step by step.

Furthermore, XGBoost employs several regularization techniques. It builds shallow decision trees, meaning they have only a few levels. Shallow trees are less prone to memorizing noise in the data and are more generalizable. Additionally, it adds a regularization term to its objective function, which penalizes complex models (i.e., models with deep trees or many splits). These penalties help control the size and complexity of the model, ensuring it generalizes well to unseen data.

XGBoost also uses subsampling techniques, where only a subset of the training data is used at each iteration, which helps reduce variance. Another useful feature of the model is early stopping, which halts training if the model's performance on a validation set stops improving. This ensures that the model does not continue training unnecessarily, which could lead to overfitting and wasted computation (overfitting is explained in Section 4.2.1).

Similarly to Random Forest, XGBoost also has the ability to rank feature importance. It keeps track of the number of times each feature is used to split data across all trees, which can help in understanding the influence of different features on the model's predictions.

4.1.3 Other Methods

There are various methods to perform classification tasks in Supervised Machine Learning. In this thesis, only the results of the methods mentioned in the previous section will be displayed on both training and testing datasets. However, other approaches, such as Multi-Layer Perceptrons (see Annex III), were initially considered for performance evaluation but did not achieve the same level of success as XGBoost or Random Forest. Additionally, methods like Support Vector Machines and K-nearest Neighbors were not explored further due to the time required for optimization or their comparatively poorer performance in the literature. Nevertheless, they are worth mentioning. It is important to note that K-nearest Neighbors, while often discussed in this context, does not technically belong to supervised learning but rather to the Lazy Learning category. This is because K-nearest Neighbors does not "learn" in the traditional sense. Instead, it retains all training instances and, when evaluating new data, compares them to the training set using a distance metric, typically the Euclidean distance.

4.2 Model Optimisation

Over the methodology's sections, terms such as Overfitting or Underfitting were mentioned, but not explained. Usually, these are addressed when optimising the machine

learning models using procedures such as Cross-Validation, in order to be avoided while improving the models' performance. Therefore, the following sections will address the optimisation process, starting by explaining the terms Overfitting and Underfitting (Section 4.2.1), followed by an explanation of Cross-Validation (Section 4.2.2).

4.2.1 Overfitting and Underfitting

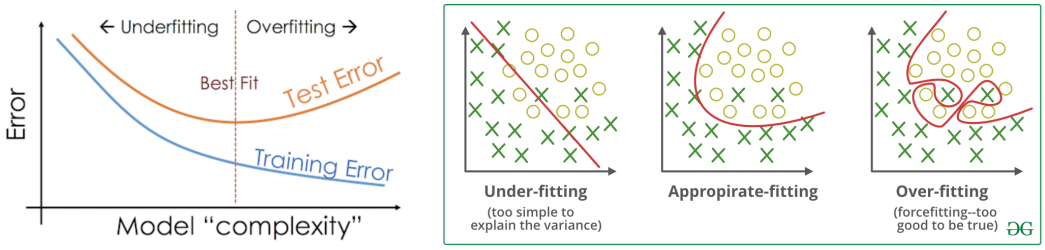
Overfitting the data refers to a model that adjusts itself too well to the training data, meaning that it learns the details and noise to an extent that negatively impacts the performance of the model in new data.

When machine learning algorithms are constructed, a sample dataset is used to train (often called training data) the model. However, when the model trains for too long on the same data or when the model is too complex, it can start to learn the "noise", or irrelevant information. Once the model memorizes the noise and fits too closely to the training set, it becomes "overfitted", being unable to generalize well new data. Consequently, if a model cannot generalize well new data, then it will not be able to perform the classification or prediction tasks that it was intended for. One way to immediately verify if a model is overfitting is by looking at the error rates and the variance; overall, low error rates and high variance are good indicators of overfitting. Another way to test that is splitting the data into training and test samples; if the training data has a low error rate and the test data has a high error rate, it is a sign of overfitting [26, 68]. Figure 4.2 (a) illustrates overfitting when the model is too complex while (b) illustrates what happens with the model in our data when it overfits.

Unlike overfitting, underfitting the data refers to a model that can neither model the training data nor generalize new data. It is easy to detect if a model is underfitting as it will have a poor performance in the training data, so it is not so often discussed.

Underfitting is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data. It occurs when a model is too simple, which can be a result of a model needing more training time, more input features, or less regularization (techniques that are used to calibrate machine learning models). Like overfitting, when a model is underfitted, it cannot establish the dominant trend within the data, resulting in training errors and poor performance of the model. The generalization of a model to new data is ultimately what allows us to use machine learning algorithms every day to make predictions and classify data.

High bias and low variance are good indicators of underfitting and since this behaviour can be seen while using the training dataset, under-fitted models are usually easier to identify than overfitted ones [26, 68]. Figure 4.2 illustrates an example of underfitting. In (a) the relationship between the error and complexity, while in (b) how the models adequate to the data when it underfits.



(a) (b)

Figure 4.2: Example of Overfitting and Underfitting - (a) relatively to the model complexity (b) relatively to the data. Figures adapted from *analyticsvidhya.com* and *geeksforgeeks.org*, respectively.

4.2.2 Cross-Validation

Cross-validation is a procedure that is used to monitor the model performance and is often used to find which hyperparameters (external configuration variables used to manage the machine learning models, e.g., number of trees in a Random Forest) should be defined when completing the model. The process involves splitting the training data into K-Folds, ensuring that each fold maintains an equal distribution of targets (stratified). During each iteration, K-1 folds are used for training the model, while one fold serves as the validation set. This procedure is repeated for a total of K iterations, where different folds are used for training and validation. Additionally, this process is repeated n times to test various hyperparameters, which by using all observations in both the training and validation sets, the model can capture the most patterns in the data while minimizing noise. This ensures the model maintains low bias and variance during the optimization procedure [26, 46, 68]. Figure 4.3 illustrates an example of this procedure for a total of 5-Folds.

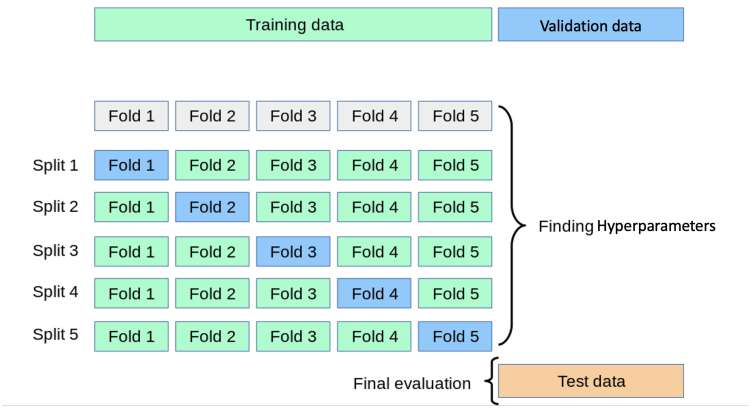


Figure 4.3: 5-Fold Cross-Validation procedure example, adapted from *scikit-learn.org*

Models' parameters configuration

Moreover, this procedure assessed different parameter combinations within the Random Forest (RF) and XGBoost (XGB) algorithms. In the case of RF, the parameters under evaluation were the number of estimators (i.e., the number of trees) and the depth (i.e., maximum splits) of the tree. For XGB, the same parameters were analysed, with an additional parameter, Gamma, included. Gamma introduces regularization within the model, which increases as the depth of the trees escalates, thereby avoiding overfitting tendencies.

It is noteworthy that the default parameters were used as the baseline for comparison. Specifically, for RF, these consisted of 100 estimators with no depth restriction (None), while for XGB, the default parameters included a depth of 6, 100 estimators, and a Gamma value of 0.

4.3 Model Evaluation

After implementing the algorithms to identify the fishing gears it is necessary to analyse their performance. As we are working with supervised machine learning it is easier to evaluate the models as we can compare the ground truth with the models' response. One way to do this would be to compare if the predictions are equal to the ground truth, however, if there is a discrepancy between the target observations, e.g., a dataset consisting of 90% of the observations with a target A and 10% of the observations with a target B, it would be easy for a model to just determine that all observations are target A, and it would achieve an accuracy of 90%, which is the times that the algorithm was coherent with the ground truth. Thus to avoid misleading conclusions, several metrics should be used to evaluate machine learning algorithms' performance. These metrics can be calculated by taking insights of the Confusion Matrix (CM) illustrated in Table 4.1. The rows of this table, nominated as T. Class x, are the True observations, while the columns, nominated as P. Class y, are the Predicted observations.

Table 4.1: Confusion Matrix for a binary classification problem

	P. Class 1	P. Class 2
T. Class 1	TP	FN
T. Class 2	FP	TN

To take insights from the CM above, we need to consider the following interpretations: True Positives (TP) are the observations from the positive class that were well classified by the model; True Negatives (TN) are the observations from the negative class that were well classified by the model. False Positives (FP), are the observations from the negative class that were classified as the positive class; False Negatives (FN), are the observations

from the positive class that were classified as the negative class. These are further used to calculate the following metrics.

- **Sensitivity** or **Recall**: The fraction of observations from the class to be analysed were correctly classified.
$$\frac{TP}{TP + FN}$$
- **Precision**: The fraction of observations classified as the class to be analysed and correctly classified.
$$\frac{TP}{TP + FP}$$
- **F1**: It is a measure that combines both precision and recall, known as a harmonized mean between both metrics.
$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$
- **Accuracy**: The fraction of observations, independently of the class, that were correctly classified:
$$\frac{TP + TN}{TP + FP + FN + TN} = \frac{Tr(CM)}{\#Observations}$$
where $Tr(CM) = \sum_{i=1}^{n} a_{ii}$, which is the sum of the main diagonal (Trace) of the confusion matrix.

These metrics are easier to calculate when estimating it for a binary problem. However, when working with a multiclass problem it may be more confusing. The procedure in this situation is to consider one of the classes as positive and the others as negative. To have a better understanding, let's consider the explanations above and evaluate Class 2 using the example in Table 4.2 to calculate some of the metrics.

Table 4.2: Illustrative example of a Confusion Matrix for multi-class classification

	P. Class 1	P. Class 2	P. Class 3
T. Class 1	12	3	2
T. Class 2	4	8	1
T. Class 3	8	5	7

As explained the accuracy is the sum of the main diagonal divided by all the observations of the dataset, in this case, the accuracy is 0.54. To evaluate the recall, let's define the TP and FN. Locking our sight into Class 2, the True Positives (TP) of the class will be the observations that were well classified by the model as Class 2, in this case, just looking at the diagonal we can estimate that it is 8. Remember that False Negatives (FN) are the observations from the positive class that were incorrectly classified as the negative class

(which includes Class 1 and Class 3). In this context, 4 observations from Class 2 were misclassified as Class 1, and 1 observation from Class 2 was misclassified as Class 3. Thus the Recall is $8/(8+(4+1)) = 0.6154$. To estimate the Precision we can follow the same logic. We already know how to estimate the TP and the False Positives (FP) are observations from the negative classes (Class 1 and Class 3) that were incorrectly classified as the positive class (Class 2). Specifically, 3 observations from Class 1 and 5 observations from Class 3 were misclassified as Class 2. Therefore, the precision is calculated as $8/(8 + (3 + 5)) = 0.5$. Lastly, the F1 score can be easily calculated after estimating the aforementioned metrics: $2 \times (0.5 \times 0.6154)/(0.5 + 0.6154) = 0.5517$.

4.4 Shapley Additive Explanations

SHapley ***A***dditive ex***P***lanations (SHAP) were introduced by Scott M. Lundberg and Su-In Lee in 2017 [35] intending to provide an interpretation of the decision making of black-box models, such as the ones mentioned in the chapters 4.1.1 and 4.1.2. The interpretation of these models is particularly challenging in cases with high complexity such as neural networks, or even when a Random Forest has multiple or deeper trees, making it difficult to explain their predictive process.

SHAP explains the prediction of an instance x by computing the contribution of each feature to the prediction, using Shapley values from the coalitional game theory proposed by L. S. Shapley, 1953 [61].

4.4.1 Shapley Values

Introduction

The Shapley value is used to explain the contribution of each feature in a predictive model for a specific instance x. It quantifies these contributions by considering all possible combinations of features. Let S represent all possible combinations of features excluding feature j. For any feature j, its contribution is given by the following formula ([42]):

$$\phi_j(val_x) = \sum_{s \subseteq S} \frac{|s|!(p - |s| - 1)!}{p!}(val_x(s \cup \{j\}) - val_x\{s\}) \tag{4.4}$$

where s is a subset of features from S, p is the total number of features, $val\{s\}$ represents the value function applied to the set of features s, and $val(s \cup \{j\})$ represents the value function applied to the set of features s plus j.

Multi-class adaptation

Despite explaining how shapley values can be calculated, it is important to highlight that in the context of the thesis these formulas can be adapted to understand the contribution

of each feature to each gear that we are predicting. Thus, the following equation 4.5 is an adaptation from equation 4.4 for a set of classes K.

$$\phi_j^k(val_x) = \sum_{s \subseteq S} \frac{|s|!(p - |s| - 1)!}{p!}(val_x^k(s \cup \{j\}) - val_x^k\{s\}) \tag{4.5}$$

where $val_x^k(s \cup \{j\})$ and $val_x^k\{s\}$ are the model's predicted probabilities for class k given the feature subsets $s \cup \{j\}$ and s, respectively.

Furthermore, $val_x\{s\}$ is calculated by (1) Sampling-Based Marginalization or (2) Analytical Marginalization (Using Expected Values). In the Sampling-Based Marginalization, we generate multiple samples for the features not in S from their empirical distribution in the training data, whereas in the Analytical Marginalization, we compute the expected values of the features not in S.

To better understand these steps, let's assume a Random Forest classifier with 3 features Season (SN), Total weight in kg caught (TKG), and length of the boat (LOA). Furthermore, we have 2 classes DRB and FPO. Defining that we would like to estimate the contribution of the feature SN for class DRB, firstly we establish all possible combinations for set S, which are {}, {TKG}, {LOA}, and {TKG, LOA}.

Considering that we are using Sampling-Based Marginalization, we generate 100 samples for each example to obtain a more accurate estimate of $\nabla(S)$. This allows us to approximate $val_x^k(S \cup \{j\})$ - $val_x^k\{S\}$ for each set as follows:

$$\begin{aligned}
\nabla(\{\}) &= val_x^{DRB}(\{SN\}) - val_x^{DRB}(\{\}) \\
&= \frac{1}{100}\sum_{i=1}^{100} \hat{f}(SN, TKG_i, LOA_i) - \frac{1}{100}\sum_{i=1}^{100} \hat{f}(SN_i, TKG_i, LOA_i)
\end{aligned}$$

$$\begin{aligned}
\nabla(\{TKG\}) &= val_x^{DRB}(\{SN, TKG\}) - val_x^{DRB}(\{TKG\}) \\
&= \frac{1}{100}\sum_{i=1}^{100} \hat{f}(SN, TKG, LOA_i) - \frac{1}{100}\sum_{i=1}^{100} \hat{f}(SN_i, TKG, LOA_i)
\end{aligned}$$

$$\begin{aligned}
\nabla(\{LOA\}) &= val_x^{DRB}(\{SN, LOA\}) - val_x^{DRB}(\{LOA\}) \\
&= \frac{1}{100}\sum_{i=1}^{100} \hat{f}(SN, TKG_i, LOA) - \frac{1}{100}\sum_{i=1}^{100} \hat{f}(SN_i, TKG_i, LOA)
\end{aligned}$$

$$\begin{aligned}
\nabla(\{TKG, LOA\}) &= val_x^{DRB}(\{SN, TKG, LOA\}) - val_x^{DRB}(\{TKG, LOA\}) \\
&= \hat{f}(SN, TKG, LOA) - \frac{1}{100}\sum_{i=1}^{100} \hat{f}(SN_i, TKG, LOA)
\end{aligned}$$

where $\hat{f}$ is the Random Forest prediction (probability) for class k (in this case DRB), SN_i, TKG_i, LOA_i are generated samples and SN, TKG, LOA are the original values.

Lastly, we can estimate the contribution of the feature Season for the class DRB as:

$$\phi_{SN}^{DRB}(val_x) = \frac{0!(2)!}{3!}\nabla(\{\}) + \frac{1!(1)!}{3!}(\nabla(\{TKG\}) + \nabla(\{LOA\})) + \frac{2!(0)!}{3!}\nabla(\{TKG, LOA\})$$

4.5 Logistic Regression

In the present work, Logistic Regression (LR) is used to interpret how the variables chosen affect the decision-making process, similar to Shapley Additive Explanations (SHAP). To do this, the LR model is trained using the training set, and its performance is assessed to determine whether or not the results are reliable for interpreting the gears' relationship with the selected variables. Nonetheless, an explanation of how LR works is provided below.

Logistic Regression is a statistical technique used to determine the likelihood of an event happening, based on the given input data. To do this, LR uses a special mathematical function called the sigmoid function to carry out this calculation, which is illustrated in Equation 4.6.

$$\theta(x) = \frac{1}{1 + e^{-x}} \tag{4.6}$$

The sigmoid function takes the input data x and transforms it into values ranging from 0 to 1. These transformed values represent the odds of the event occurring. For instance, if the sigmoid function gives a vector of odds and one of the values is 0.8, for a particular trip, it means there is an 80% chance that the boat fished with that specific gear.

Similar to a linear regression, through this technique we can model y to x as follows:

$$y = \frac{1}{1 + e^{-(\beta_0 + wX)}} \tag{4.7}$$

where β_0 is the intercept w is the vector of the parameters to be optimised, which corresponds to the number of features included in the data and X our input data of shape (n, m), with n being the number of observations and m the number of features.

Therefore, after training a LR model it will provide us with coefficients relative to each feature, and thus, it is possible to make conclusions about the features and the respective gear. However, when there are multiple outputs, such as the current work (eight gears), to perform a multinomial Logistic Regression, two common approaches can be used: One-Vs-Rest (OvR) or Multinomial (uses Softmax function). In OvR methodology, LR fits K binary classifiers (one for each class) and each classifier k is trained to distinguish class k from all other classes. Thus, providing K sets of coefficients per class. The final decision with OvR will be the classifier with the highest probability among the K classifiers. On the other hand, the Multinomial fits a single model with K-1 sets of coefficients, since the K-th set of coefficients can be derived from the others due to the normalization constraint of the softmax function (see Equation 4.8). Hence, the K-th class probability is implicitly

determined by the others. Regardless, Scikit-learn, the python library used in the current work, provides the K set of coefficients independently of the approach used.

$$P(y = k \mid x) = \frac{e^{(\beta_k + w_k X)}}{\sum_{j=1}^{K} e^{(\beta_j + w_j X)}} \tag{4.8}$$

After building the model it is possible to do inference on the model's coefficients, i.e., test if a particular coefficient is statistically different from zero, which helps determine if the predictor variable has a significant impact on the outcome. Therefore, to perform this analysis we can test the following hypothesis (Equation 4.9).

$$H_0 : \ \beta_i = 0 \ \ vs \ \ H_1 : \ \beta_i \neq 0 \tag{4.9}$$

where β_i is the coefficient related to the i-th feature.

Furthermore, the Central Limit Theorem was considered, which justifies that the distribution of the sample coefficients (estimates) approximates a normal distribution as the sample size grows. Additionally, since the Maximum Likelihood Estimation (see [14] for more details) was used to calculate these estimates, we can assume the estimates' asymptotic normality.

Therefore, to attest the hypothesis, by using a stratified bootstrap strategy (resampling the original data) with replacement, the Z-score was calculated as follows:

$$z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

where $\hat{\beta}_i$ is the mean of the estimates and $SE(\hat{\beta}_i)$ is the standard error of the estimates.

Considering the normal distribution we can further calculate the P-value as follows:

$$p = 2 \times (1 - \Phi(|z|))$$

where Φ is the cumulative distribution function of the standard normal distribution.

Finally, if the p-value is lower than a specific threshold (significant level), usually $\alpha = 0.05$, we reject the null hypothesis H_0, which means that feature β_i hold a significant impact on the model's predict process.

5

Results

This section presents the results of the thesis. Two classification approaches were taken: (1) classifying all gear types individually, and (2) classifying all gear types but combining Gillnets and Trammel Nets into a single category. Results will be shown without discussion, which will be covered in the Discussion section (6).

First, the outcomes of the training process and hyperparameter tuning will be presented. Next, the models, using the best hyperparameters, were tested on unseen data, with a detailed analysis of feature importance, focusing primarily on XGBoost, which emerged as the best-performing model. XGBoost was also applied to the Portuguese dataset. Finally, the relationships between features and gear types are explored using Shapley Additive Explanations (SHAP) and Logistic Regression (LR).

5.1 Parameters Selection and Training Performance

Approach 1: Classification of all gears

Table 5.1 illustrates the average accuracy score in the 5-fold Cross-Validation in Random Forest, where the lowest score obtained was 89.25% in the validation set using 100 estimators and a depth of 5. The best parameters combination of parameters was 100 estimators and the default depth (None) achieving an average accuracy of 94.50% on the validation set. The accuracy of both train and validation sets was similar not demonstrating evidence of overfitting.

Table 5.1: Random Forest 5-Fold Cross-Validation procedure average performance - all gears

Estimators	Depth	Train Score	Val Score
100	None	98.27	94.50
	5	89.38	89.25
	10	93.41	92.63
	15	96.18	93.58

150	None	98.27	94.39
	5	89.43	89.39
	10	93.36	92.58
	15	96.21	93.59
200	None	98.28	94.43
	5	89.50	89.43
	10	93.42	92.59
	15	96.22	93.54
300	None	98.28	94.45
	5	89.39	89.32
	10	93.46	92.70
	15	96.22	93.57

On the other hand, Table 5.2 illustrates the performance of XGBoost through the 5-fold Cross-Validation procedure. Within the combinations tested, the lowest score obtained was 94.54% using 200 estimators, a depth of 15, and a Gamma equals 0.2. As for the best combination, a validation score of 94.83% was obtained with Gamma equals 0.01, a depth of 5, and 100 estimators. Similarly to Random Forest, XGB obtained a similar score in both train and validation sets, not showing any evidence of overfitting.

Table 5.2: XGBoost 5-Fold Cross-Validation procedure average performance - all gears

Estimators	Depth	Gamma	Train Score	Val Score
100	5	0.01	97.86	94.83
	6	0	98.10	94.70
	10	0.1	98.27	94.68
	15	0.2	98.26	94.55
150	5	0.01	98.12	94.69
	6	0	98.24	94.66
	10	0.1	98.27	94.61
	15	0.2	98.25	94.65
200	5	0.01	98.23	94.73
	6	0	98.28	94.62
	10	0.1	98.29	94.68
	15	0.2	98.26	94.54
300	5	0.01	98.29	94.55
	6	0	98.28	94.71
	10	0.1	98.28	94.63
	15	0.2	98.26	94.55

Approach 2: Aggregating the nets

Table 5.1 illustrates the average accuracy score in the 5-fold Cross-Validation in Random Forest, where the lowest score obtained was 93.68% in the validation set using 100 estimators and a depth of 5. The best parameters combination of parameters was 150 estimators and the default depth (None) achieving an average accuracy of 99.13% on the validation set. The accuracy of both train and validation sets was similar not demonstrating evidence of overfitting.

Table 5.3: Random Forest 5-Fold Cross-Validation procedure average performance - Nets aggregated

Estimators	Depth	Train Score	Val Score
100	None	99.99	99.02
	5	93.81	93.68
	10	97.25	96.97
	15	98.24	97.81
150	None	100.0	99.13
	5	93.81	93.72
	10	97.14	96.88
	15	98.25	97.79
200	None	99.99	99.13
	5	93.92	93.86
	10	97.19	96.95
	15	98.23	97.79
300	None	99.99	99.13
	5	93.87	93.80
	10	97.18	96.86
	15	98.23	97.79

On the other hand, Table 5.2 illustrates the performance of XGBoost through the 5-fold Cross-Validation procedure. Within the combinations tested, the lowest score obtained was 99.15% using 100 estimators, a depth of 5, and a Gamma equals 0.01. As for the best combination, a validation score of 99.40% was obtained with Gamma equals 0.01, a depth of 5, and 300 estimators. Similarly to Random Forest, XGB obtained a similar score in both train and validation sets, not showing any evidence of overfitting.

Table 5.4: XGBoost 5-Fold Cross-Validation procedure average performance - Nets aggregated

Estimators	Depth	Gamma	Train Score	Val Score
	5	0.01	99.84	99.15

100	6	0	99.94	99.30
	10	0.1	99.99	99.31
	15	0.2	99.95	99.22
150	5	0.01	99.95	99.25
	6	0	99.99	99.29
	10	0.1	99.99	99.34
	15	0.2	99.96	99.23
200	5	0.01	99.99	99.33
	6	0	99.99	99.35
	10	0.1	99.99	99.26
	15	0.2	99.96	99.24
300	5	0.01	99.99	99.40
	6	0	99.99	99.36
	10	0.1	99.98	99.26
	15	0.2	99.97	99.24

5.2 Performance on unseen data

Approach 1: Classification of all gears

Random Forest (RF) and XGBoost (XGB) achieved similar performances using the variables kg per species, number of species caught, month of the landing, LOA, and the total kg caught, with 94% accuracy (Table III.1), although the latest outperforms by 7% RF when detecting GNS. All other gears have similar performances between models, showing similar average precision and recall. Different combinations of variables were tested, (e.g. using only total kg per species and month, adding LOA, etc), but slightly lower accuracies were obtained (around 91% accuracy)(not shown for brevity). Further, instead of using the taxa groups, the model considering the raw species also showed the worst accuracies (not shown for brevity). Further models, using presence absence and percentage caught by trip were tested for all gears and only for the worst cases (GNS/GTR) but gave similar results and thus were not used for the sake of simplicity.

Table 5.5: RF and XGB performances, when using the taxa caught in kg, taxa richness (S), LOA, total kg caught by trip (independently of the taxa) and month of the landing to identify gears within IEO landing data.

Gear / Model	RF	XGB
	F1-Score	
DRB	100%	100%
FPO	100%	100%

GNS	44%	51%
GTR	91%	92%
LHP	93%	90%
LLS	96%	94%
OTB	99%	99%
PS	98%	99%
Accuracy	94%	94%
Avg. Precision	91%	91%
Avg. Recall	90%	90%

The higher misclassification rate observed for both RF and XGB, was between GNS and GTR, with GNS showing a poorer performance among all gears (Figure 5.1). There were also some misclassifications between LHP and FPO whereas some LHP landings are confused with FPO (9.52% RF and 15.08% XGB). All the other gears were well classified showing scores above 90%, in particular, both DRB and FPO had scores of 100% and 99.86%, respectively.

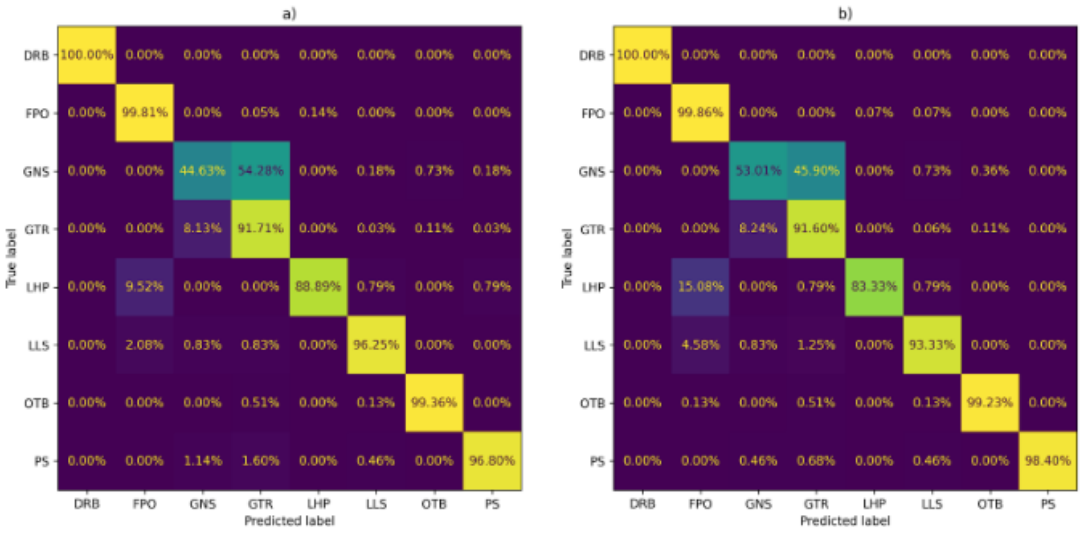


Figure 5.1: Confusion Matrix a) Random Forest and b) XGBoost

Approach 2: Aggregating the nets

Similarly to the previous approach, both models achieved similar performances, where RF and XGB achieved an accuracy and precision of 99%. In terms of recall, XGB slightly outperformed RF (Table 5.6). Within the gears, LHP and LLS were the ones that were more misclassified, and aggregating both GNS and GTR (Nets) as one class considerably improved the classification of these gears.

Table 5.6: RF and XGB performances, when using the taxa caught in kg, taxa richness (S), LOA, total kg caught by trip (independently of the taxa) and month of the landing to identify gears (with GNS and GTR as one class) within IEO landing data.

Gear / Model	RF	XGB
	F1-Score	
DRB	100%	100%
FPO	100%	100%
Nets	100%	100%
LHP	91%	93%
LLS	96%	94%
OTB	100%	99%
PS	98%	99%
Accuracy	99%	99%
Avg. Precision	99%	99%
Avg. Recall	96%	97%

The higher misclassification rate observed for both RF and XGB, was between LHP and FPO, with LHP showing a poorer performance among all gears (Figure 5.1). All the other gears were well classified showing scores above 90%, including the new aggregated class of both Nets (Gillnets + Trammel Nets).

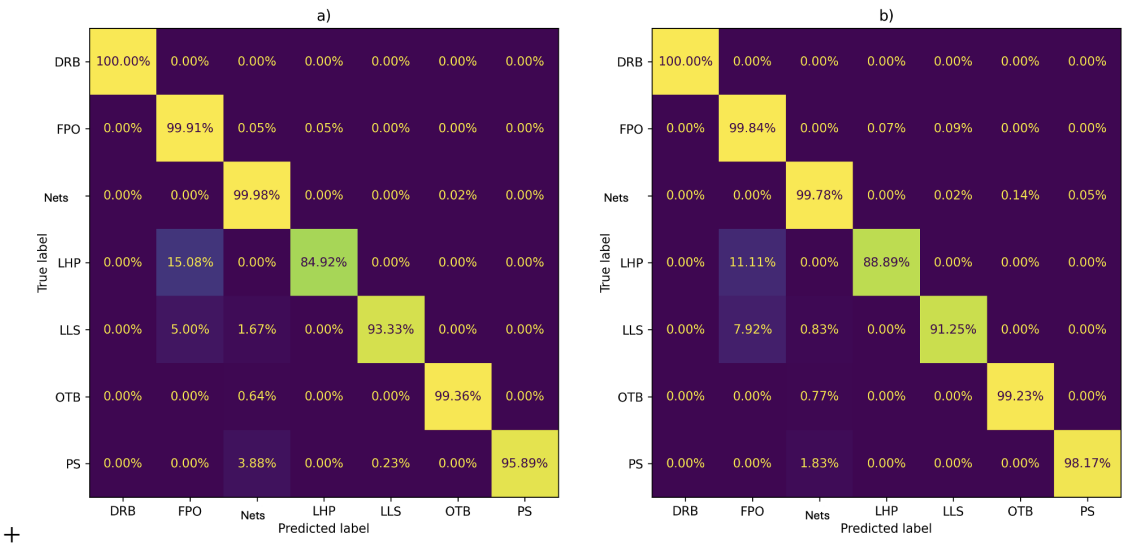
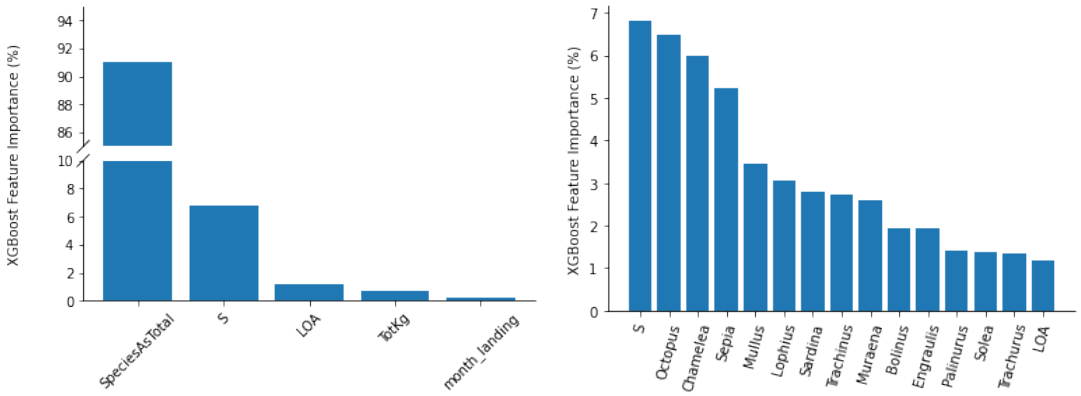


Figure 5.2: Confusion Matrix a) Random Forest and b) XGBoost

5.3 Features predictive power

The most important variable used in the ML classification was the taxa caught in kg (91.42%), followed by the taxa richness (6.82%), LOA (1.12%), total kg caught by trip (independently of the taxa) (0.73%) and month of the landing (0.26%)(Figure 5.3 a). The five most important taxa were Donax (38.99%)(not present on the visualization to have a better preview of the other variables), Octopus (6.48%), Chamelea (5.98%), Sepia (5.21%) and Mullus (3.44%)(Figure 5.3 b). As for the least important features, seven taxa had a zero-importance contribution to the model - Dipturus, Spicara, Brama, Torpedo, Leucoraja, Lepidotrigla, and Myliobatis.



(a) (b)

Figure 5.3: XGBoost feature importance (a) Overall Feature Importance and (b) Individual Feature Importance

5.4 Classification of the Portuguese Landings

Approach 1: Classification of all gears

Once the best model was developed for the IEO dataset (XGBoost using as explicative variables taxa caught (kg), S, LOA, total kg caught by trip, and month of the landing), it was then applied to the dataset from Algarve (Portugal, IPMA). Table 5.7, shows the results of applying the model in the IPMA dataset. The Main | Sub. column indicates the proportion of the classification by the model that matched the primary or subsidiary gear licences (in Portugal boats can have several gears authorised). In this case, DRB, FPO, GTR, OTB, and PS had higher matches with the model (95.16%, 76.38%, 75.41%, 79.25%, and 80.25%, respectively). A lower percentage of matches was obtained for the LHP, LLS, and GNS gears (48.34%, 52.28%, and 67.74%).

Table 5.7: Results by applying XGB onto IPMA landings' dataset. It compares the model classification with the main gear licenses and the % of matches between the main and subsidiary gears.

		Model Classification							
Main Gear	**Main | Sub.**	**DRB**	**FPO**	**GNS**	**GTR**	**LHP**	**LLS**	**OTB**	**PS**
DRB	95.16	90.63	8.79	0.23	0.34	0.00	0.00	0.00	0.00
FPO	76.38	8.88	59.58	8.78	20.53	0.21	1.09	0.07	0.87
GNS	67.71	1.78	34.78	12.70	47.61	0.23	1.48	0.09	1.33
GTR	75.41	1.18	27.11	13.33	56.74	0.18	0.93	0.15	0.38
LHP	48.34	5.92	35.39	19.44	37.32	0.50	1.43	0.01	0.00
LLS	52.28	10.11	38.54	12.07	37.28	0.00	1.52	0.48	0.00
OTB	79.25	0.00	8.49	0.94	11.32	0.00	0.00	79.25	0.00
PS	80.25	4.75	11.02	9.88	23.27	0.00	0.09	0.00	51.00

Approach 2: Aggregating the nets

For the second approach XGBoost using as explicative variables taxa caught (kg), S, LOA, total kg caught by trip, and the month of the landing was too the best model developed. Hence, it was then applied to the dataset from Algarve (Portugal, IPMA). Table 5.8, illustrates the results of applying it, whereas the Main | Sub. column indicates the proportion of the classification by the model that matched the primary or subsidiary gear licences. In this case, DRB, FPO, Nets (Gillnets + Tramell Nets), OTB, and PS had higher matches with the model (95.33%, 87.75%, 81.49%, 84.91%, and 91.74%, respectively). A lower percentage of matches was obtained for the LHP and LLS gears (53.98% and 70.00%, respectively).

Table 5.8: Results by applying XGB onto IPMA landings' dataset. It compares the model classification with the main gear licenses and the % of matches between the main and subsidiary gears, considering Gillnets and Trammel Nets aggregated as a single class

		Model Classification						
Main Gear	**Main | Sub.**	**DRB**	**FPO**	**Nets**	**LHP**	**LLS**	**OTB**	**PS**
DRB	95.33	90.63	8.79	0.57	0.00	0.00	0.00	0.00
FPO	87.75	8.81	59.68	29.40	0.23	0.93	0.09	0.86
Nets	81.49	1.70	35.29	60.23	0.16	1.07	0.20	1.35
LHP	53.98	5.88	35.27	57.06	0.48	1.29	0.01	0.00
LLS	70.00	10.07	38.59	49.56	0.02	1.26	0.50	0.00
OTB	84.91	0.00	8.49	6.60	0.00	0.00	84.91	0.00

PS	91.74	4.75	11.02	32.29	0.00	0.09	0.28	51.57

Further, the results of the model and the license information were compared with expert validation whereas the model implemented was coherent with the expert validation with 82.6% accuracy and the main and subsidiary gear with 81.23%.

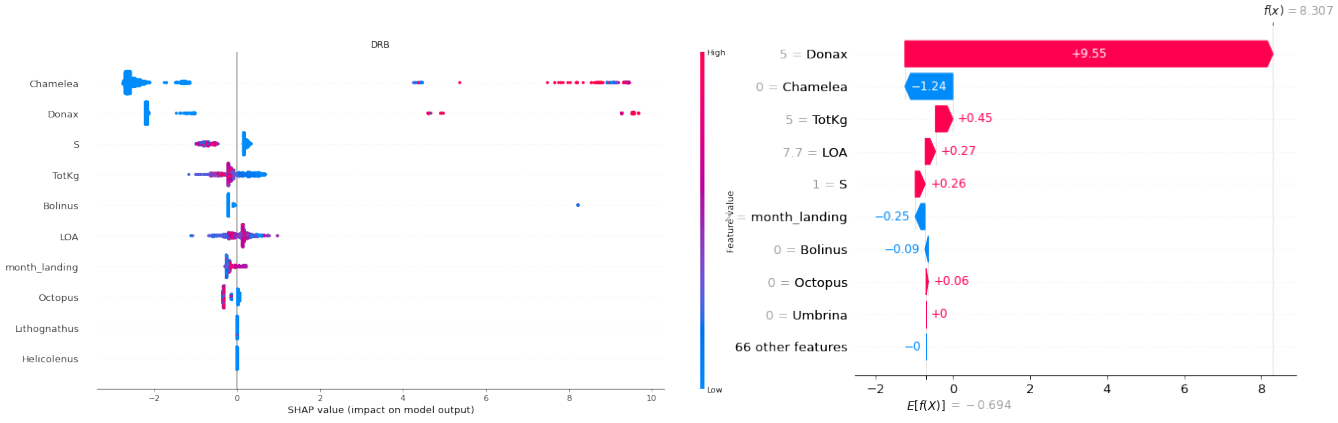
5.5 Relationship between the gears and the features

Explained by Shapley Additive Explanations

Shapley Additive Explanations (SHAP) illustrate the most impactful variables to classify each gear. In the following figures, only the ten most important features were illustrated for each gear. Moreover, in all figures a) the right side of the figure indicates positive contributions and the left negative contributions. The colour shows if that contribution happens with feature high values (in red) or lower values (in blue). Additionally, the figures on the right (the figures b) illustrate the individual contribution for a randomly selected trip for the specific gear being analysed.

Given this information, DRB is strongly related to catching high amounts of Chamelea and Donax (Figure 5.4 a). Further, there is no other feature that has such an impact as the previous ones. There is a small percentage of influence if more species are caught within these fisheries, which contributes negatively to the classification as DRB.

Furthermore, it is possible to analyse the model decision-making for an individual trip. Figure 5.4 b) illustrates that catching Donax highly contributes for the model to classify this trip as Dredges as the Shapley value is greater than 8.

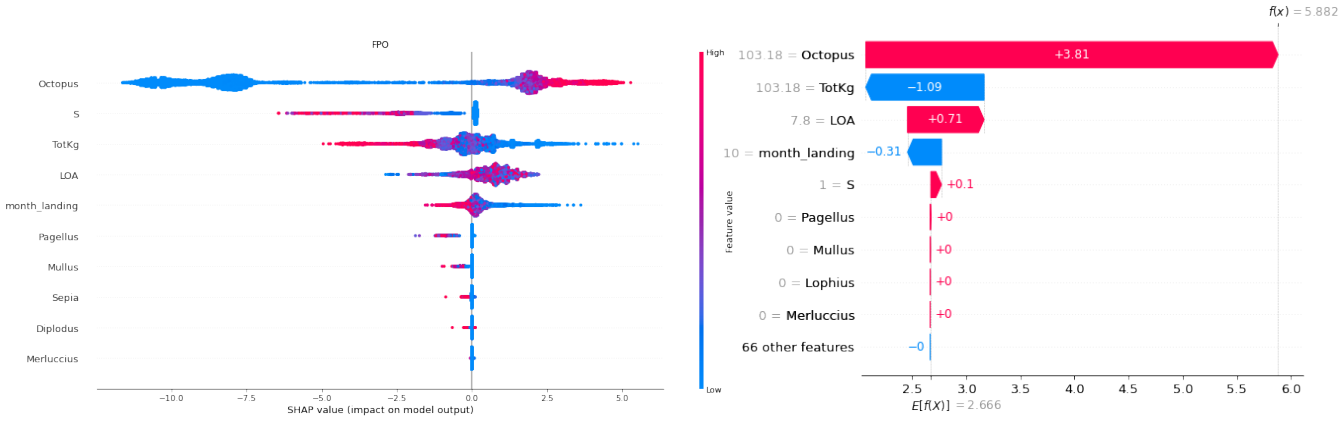


(a) (b)

Figure 5.4: Feature Importance explained by SHAP within fisheries with DRB - (a) Globally and (b) Individually for a randomly selected trip

The results of SHAP show higher values of Octopus caught contribute to a positive influence of classifying the landings as FPO, and the presence of low catches would

indicate that the gear is not FPO. The second more important value would be the number of species caught, where the presence of a high number of species by trip (S) indicates that the gear is not FPO and a low number of species also contributes positively to the identification, confirming the species-specific nature of this fishery. The aggregation of more red dots on the right side of the figure shows that boats with an LOA close to the average tend to operate with this gear. On the other hand, the total kg caught and the month of the landing contribute negatively to the classification of FPO boats, meaning that there are higher catch rates in the initial months of the year and that usually this fishery does not deliver high catches (Figure 5.5 a). Additionally, Figure 5.5 b) shows similar outputs as the previous figure, illustrating for an individual trip that catching Octopus contributes positively to the classification of the gear. Nonetheless, having many catches in kilograms negatively influences the classification.

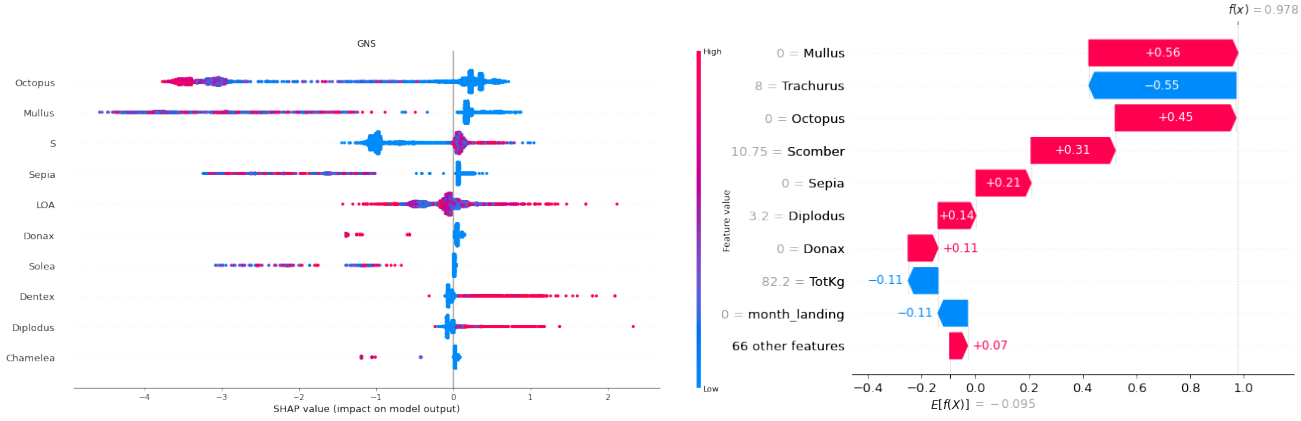


(a) (b)

Figure 5.5: Feature Importance explained by SHAP within fisheries with FPO - (a) Globally and (b) Individually for a randomly selected trip

On the other hand, the results for GNS showed that higher catches of Diplodus and Dentex positively contribute to the gear classification. Moreover, high catches of Octopus, Solea, Sepia and Mullus contribute negatively to classifying a landing as GNS. Lastly, having a lower variety of species caught (taxa richness) negatively influences the classification (Figure 5.6 a). The decision-making process for an individual trip is illustrated in Figure 5.6 b), whereas Scomber and Diplodus were the species that were caught and most contributed to the gear classification.

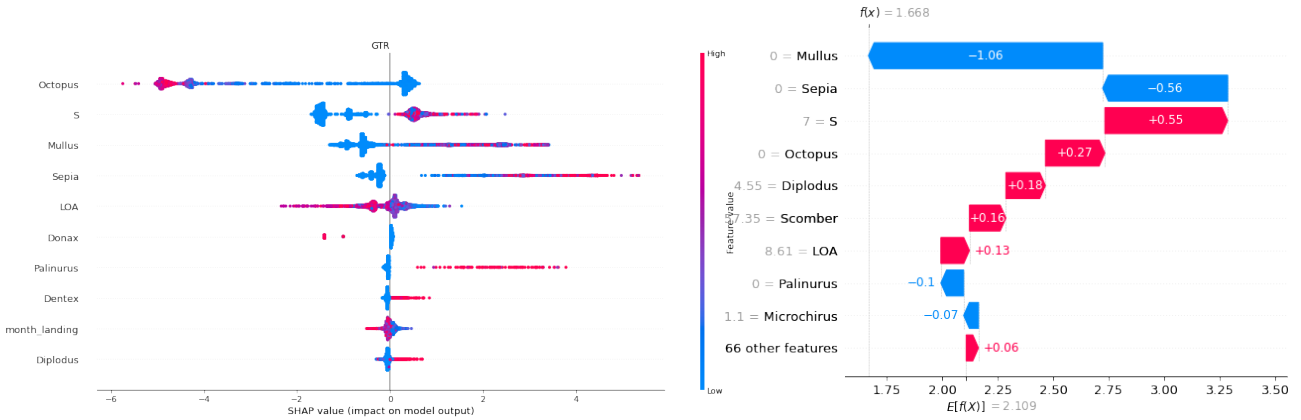
The boats having larger LOA have a negative influence on classifying GTR. Mullus, Sepia, and Palinurus were found to contribute positively to the gear classification. Furthermore, higher catches of Diplodus and Dentex tend to influence a landing to be considered as GTR. Also, as more species are caught there is a positive impact on the classification of this gear (Figure 5.7 a). Moreover, by analysing an individual trip of GTR it is possible to conclude that having a greater taxa richness, cathing Diplodus and Scomber contributed to the gear classification while not having caught Mullus highly influenced the model to



(a) (b)

Figure 5.6: Feature Importance explained by SHAP within fisheries with GNS - (a) Globally and (b) Individually for a randomly selected trip

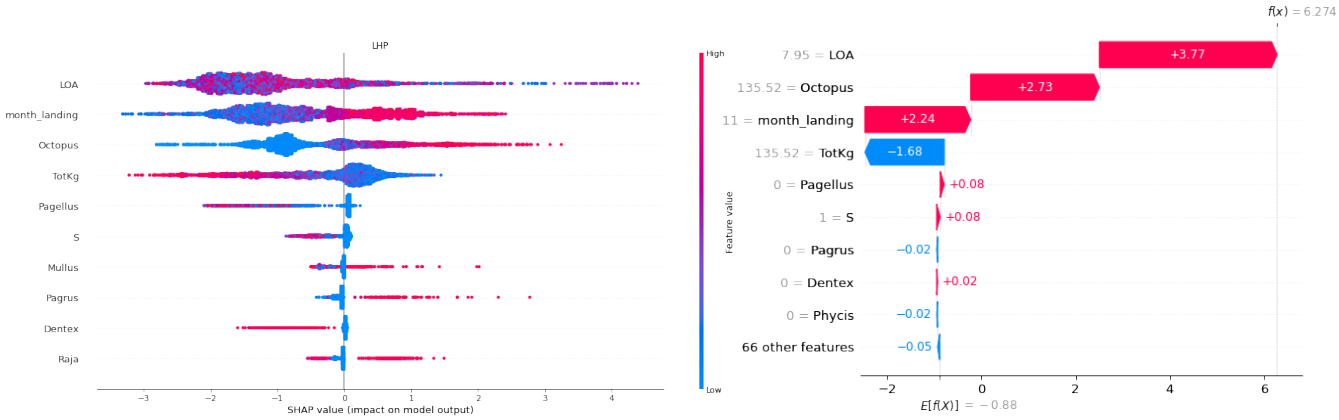
not select this gear as the final classification (Figure 5.7 b).



(a) (b)

Figure 5.7: Feature Importance explained by SHAP within fisheries with GTR - (a) Globally and (b) Individually for a randomly selected trip

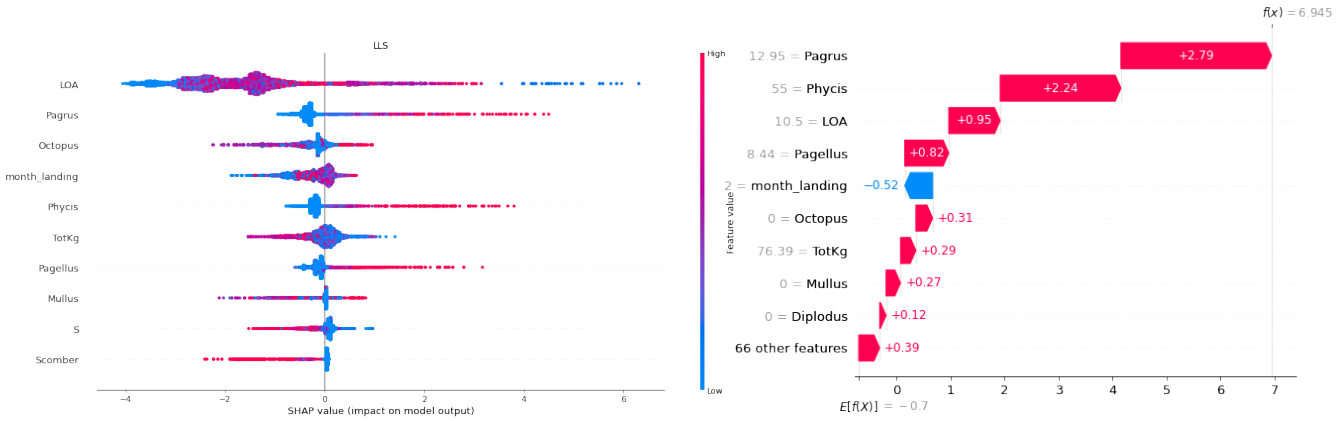
High quantities of Mullus and Pagrus characterise LHP. Further, these fisheries are related to more activity at the end of the year as high values for the month variable contribute positively to the classification of LHP. Moreover, Octopus has a positive influence with higher catches and a negative influence with lower catches. Overall, higher quantities of Pagellus, Dentex, and a higher number of different species caught in the same boat trips, illustrate a negative contribution to a landing to be classified as LHP (Figure 5.8 a). Nonetheless, Figure 5.8 b) illustrates some of the patterns mentioned, by having Octopus caught at the end of the year and the gear being operated on a boat with an overall length (LOA) of 7.95 contributed positively to the classification of this trip as LHP.



(a) (b)

Figure 5.8: Feature Importance explained by SHAP within fisheries with LHP - (a) Globally and (b) Individually for a randomly selected trip

LLS has high catches of Pagrus, Pagellus, and Phycis. On the other hand, Scomber and the taxa richeness influence negatively this class by having higher values (Figure 5.9 a). Moreover, in a randomly select trip of LLS, it is illustrated that of the three species Pagrus had a greater influence on the decision-making process towards the gear, followed by Phycis and Pagellus (Figure 5.9 b).

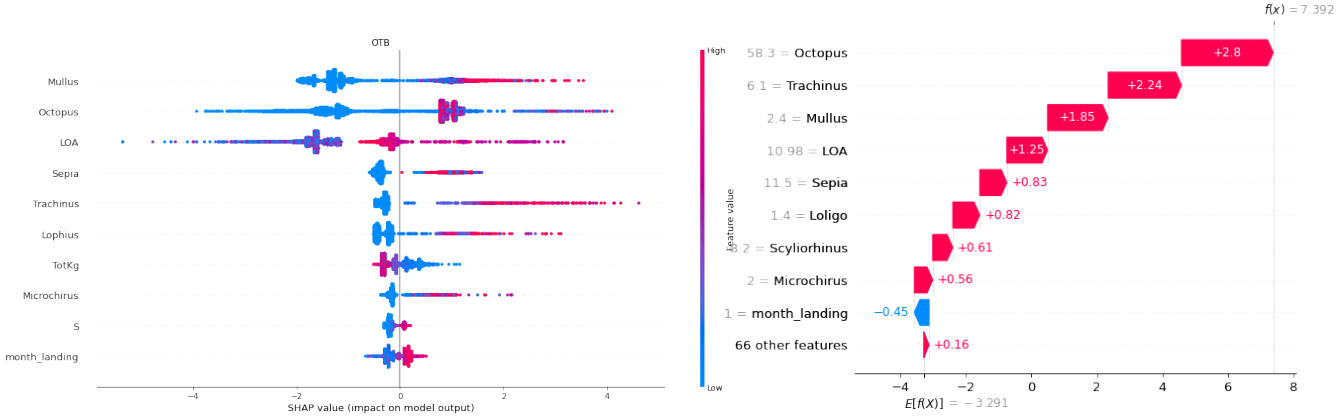


(a) (b)

Figure 5.9: Feature Importance explained by SHAP within fisheries with LLS - (a) Globally and (b) Individually for a randomly selected trip

As for OTB, Figure 5.10 a) shows that larger boats tend to operate this gear. Further, despite a lower contribution to the classification, it is shown that this gear has more activity at the end of the year. Moreover, higher catches of Octopus, Lophius, Mullus, Microchirus, Trachinus, and Sepia tend to contribute positively to the classification of the gear. Lastly, the total of resources captured tends to classify more observations as OTB when low amounts are registered. Figure 5.10 b) illustrates for a specific trip, that catching

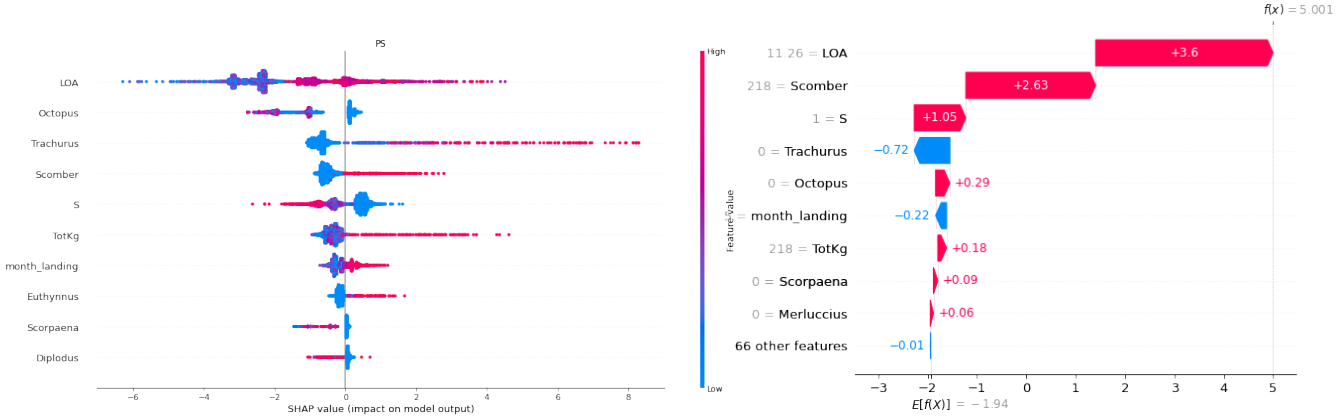
several species of the ones mentioned above, within the same trip highly contribute to the classification of this gear.



(a) (b)

Figure 5.10: Feature Importance explained by SHAP within fisheries with OTB - (a) Globally and (b) Individually for a randomly selected trip

PS tends to catch a higher quantity of resources (TotKg) and have more activity at the end of the year. Further, the species that have higher catch rates and positively influence the classification of this gear are Trachurus, Scomber, and Euthynnus. On the other hand, Scorpaena and Diplodus contribute negatively when present with higher catch rates. Additionally, a high variety of species caught has a negative influence. Lastly, larger boats tend to operate this gear (Figure 5.11 a). Furthermore, it is illustrated for a specific trip that the features that contributed more to the gear classification were a LOA of 11.26, catching a high amount of Scomber and being the only species (Figure 5.11 b).



(a) (b)

Figure 5.11: Feature Importance explained by SHAP within fisheries with PS - (a) Globally and (b) Individually for a randomly selected trip

Explained by Logistic Regression

Logistic Regression achieved a reasonable performance (93% of accuracy), but it performed poorly in precision and recall: 79% and 72%, respectively (Table 5.9). Furthermore, the gears that had poor performance were GNS, LHP, and LLS.

Table 5.9: LR performance, when using the taxa caught in kg, taxa richness (S), LOA, total kg caught by trip (independently of the taxa) and month of the landing to identify gears within IEO landing data.

Gear / Model	LR F1-Score
DRB	100%
FPO	98%
GNS	62%
GTR	94%
LHP	2%
LLS	67%
OTB	85%
PS	85%
Accuracy	93%
Avg. Precision	79%
Avg. Recall	72%

Moreover, by making inferences into the logistic regression coefficients for each gear (see Section 4.5), it was selected the coefficients with values greater than 0.50 as several variables were statistically significant (p-value < 0.05)(Table 5.9).

With that in mind, the coefficients highlighted positively by LR in Dredges (DRB) were Bolinus, Chamelea, and Donax. Relatively to Pots & Traps (FPO), Conger, Diplodus, Octopus, and Sepia were the coefficients with more positive influence. On the other hand, catching Mullus, Pagellus and having a higher amount of total catches contributed the most for a trip not being classified as FPO. Gillnets (GNS) had greater coefficients on variables that negatively influence the gear classification such as Mullus, Sepia, and Solea. The species that positively contributed more were Lophius and Merluccius. The Trammel Nets (GTR) coefficients that contributed more to the classification were Sepia, Pegusa, and Mullus. As for the negative contribution, it was highlighted by Chamelea and Donax. The LR model defined Chamelea and Mullus as the species that most contributed to Hand & Pole Lines (LHP). Additionally, having a higher taxa richness negatively contributed to the gear classification. Moreover, Longlines (LLS) had as species that contributed more to this class Muraenam and Phycis. Nevertheless, Lophius and Sepia were the species that negatively influenced the decision-making process towards this gear. Otter

Bottom Trawlers (OTB) have several variables that positively or negatively affect the gear classification. Sepia, Mullus, Solea, Lophius and the taxa richness (S) were the ones that contributed more positively, while Diplodus, Phycis, and Scomber were the ones that contributed more negatively. Lastly, the species that most contributed positively to the Pursue Seiners (PS) gear classification were Scomber, Sardina, Mullus, and Pomadasys. On the contrary, Merluccius, Octopus, and Scorpaena were the species that negatively affected the model decision-making process towards this gear.

Table 5.10: Logistic Regression coefficients with an absolute value greater than 0.5 that were statistical significant for each gear

Gear 1	Variables 1	Coeff. 1	Variables 2	Coeff. 2	Variables 3	Coeff. 3
DRB	Bolinus	1.11	Pagellus	-0.67		
	Chamelea	1.77	Sepia	-0.51		
	Donax	2.77	S	-0.57		
	Octopus	-0.72				
FPO	Conger	1.18	Pagellus	-1.08		
	Diplodus	1.20	Sepia	2.32		
	Mullus	-1.20	TotKg	-1.14		
	Octopus	1.48	Dentex	-0.67		
GNS	Chamelea	-0.70	Plectorhinchus	0.57	Solea	-1.77
	Lophius	0.72	Polyprion	-0.60	Spicara	0.53
	Merluccius	0.73	Scomber	0.52	Uranoscopus	0.66
	Mullus	-3.45	Sepia	-2.80	Zeus	0.59
GTR	Chamelea	-1.14	Sepia	1.21		
	Donax	-0.77	Solea	0.61		
	Mullus	0.81	Spicara	0.67		
	Pegusa	0.84	TotKg	0.53		
LHP	Chamelea	1.18	Muraena	0.53		
	Conger	-0.59	S	-1.16		
	Donax	-0.61				
	Mullus	0.88				
LLS	Dentex	0.58	Muraena	0.75	Seriola	-0.61
	Euthynnus	-0.63	Phycis	0.71	Serranus	0.60
	Homarus	0.60	Plectorhinchus	-0.56		
	Lophius	-1.14	Sepia	-1.20		
OTB	Balistes	-0.75	Loligo	0.72	Mullus	1.41
	Boops	-0.73	Lophius	1.08	Pagellus	0.58
	Diplodus	-1.05	Merluccius	1.03	Phycis	-0.93
	Lithognathus	-0.54	Microchirus	0.62	Scomber	-0.87
	Scorpaena	0.60	Trachinus	0.51	Sepia	1.59
	Uranoscopus	-0.58	Solea	1.25	LOA	-0.51
	Spicara	-0.80	S	1.05		
PS	Merluccius	-0.80	Sardina	0.58		
	Mullus	0.58	Scomber	0.57		
	Octopus	-0.72	Scorpaena	-0.70		
	Pomadasys	0.56				

6

Discussion

The current work focuses on identifying the fishing gears operated by the Small-Scale fleet in Portugal, Algarve. However, the information about the gears was not available for the Portuguese fleet. Thus, data from the neighbouring country Spain was used to implement a machine learning model which was evaluated and subsequently applied to the Portuguese data.

Data Preprocessing and Cleaning

The first step of this thesis was to unify the dataset from both countries by matching fishing gears and removing noisy observations, which consist of rare species and trips with very low catches. A data-cleaning process was implemented to remove species caught in less than 10 trips and trips with a total catch below 0.5 kg. Further, removing the boats with overall lengths above 12 meters and matching the genera between datasets reduced the original number of trips of the IEO dataset by 22.2% and IPMA's dataset by 10.1%.

Leitão et al. (2022) applied a similar approach to the one applied in the current thesis to remove the noisy observations within their dataset. The authors' process consisted of (1) selecting the 100 species with the highest landing frequencies (2) selecting the 50 species with the highest total landings (in kilograms) and (3) selecting the trips that landed more than 10% of the average daily landing of the specific species. By applying this step a decrease of 3.8% in the number of trips was obtained [32].

These data-cleaning procedures highlight the critical role in producing reliable and valid results. By removing irrelevant or anomalous trips, the quality of the dataset is enchanted, which directly impacts the subsequent analyses. Although the percentage of removed data varies between the current thesis and the cited work [32], the objective of constructing a representative dataset that provides a reliable foundation for meaningful analysis remains.

Feature Selection

The objective of this thesis is to identify fishing gears based on landing profiles. A key aspect of this process is understanding which species are caught by each gear, as species

information in the landing profiles is essential for accurately identifying the corresponding fishing gears. Since certain species are typically associated with specific types of gear, it is important to incorporate species-related indicators such as catch weight (in kilograms), total price per kilogram, or presence/absence of species—into the modelling process, enhancing the accuracy and reliability of gear identification [65].

Moreover, season variability has been used to understand the correlation within species. Szynaka et al (2021) illustrate that Octopus and Donax were caught mostly over the year with no specific seasonal trends, which were demonstrated to be related to Pots & Traps and Dredges respectively [65]. These insights are similar to what was demonstrated in the explanatory data analysis, whereas the two gears had fishing activity almost every month of the year. This reinforces the relationship between the species and gears and how the activity can be affected by the season variability.

Nonetheless, Shester et al. (2011) [62] have also demonstrated that the overall length of the boats (LOA) may vary and be related to the selection of some gears or the number of gears being carried (e.g. Pots).

Lastly, from the variables selected for the models, taxa richness (S) introduced a factor of diversity as several fisheries may use one gear that caught several species or just one species, as illustrated in the exploratory data analysis. Thus, this variable would be important to distinguish some gears. These findings were also highlighted in some studies reinforcing the impact this variable may have on the decision-making process of the models [5, 11, 56].

Models' Performance

Once the variables to be included in the models, i.e., species weight in kg (71 species), taxa richness, the month of the landing, the overall length of the boat, and total weight in kg caught, Random Forest (RF) and XGBoost (XGB) had excellent performances on the Spanish dataset, not demonstrating any sign of overfitting (94% accuracy on unseen data). Nevertheless, the algorithms illustrated misclassifications between Gillnets (GNS) and Trammel Nets (GTR). After discussing these results with the experts, it was concluded that even looking at the available data, it would be hard to distinguish between the two gears. This made us implement a new algorithm aggregating the two gear types, reaching an accuracy of 99% on unseen data. This new algorithm had excellent performances, yet there was a misclassification between the gear Pots & Traps and LHP (15% RF and 11% XGB). The analysis conducted concluded that both gears were fishing Octopus which led the algorithm to misclassify some of the cases, as Pots and Traps, as usually it would catch more octopus in weight and with higher frequencies.

Nonetheless, one of the objectives of this thesis was to use the models implemented in the Spanish dataset to evaluate the Portuguese landing profile data and identify the fishing gear in the Algarve (south Portugal). The procedure was similar whereas at first the model with all gears was tested achieving a coherency with the main and subsidiary

licenses of 71.85% while the model with the nets aggregated achieved a coherency of 80.74%. In both cases, it was possible to analyse that despite the higher precision of the main or subsidiary gear, the performance on the main license was always smaller. This may indicate that in some cases, especially in LLS, LHP, and Nets, these fisheries tend to operate with their subsidiary gear, which can reduce the reliability of data based solely on the main license, given their flexibility in gear usage.

Similar performances were obtained by Palmer et al. (2017) [45]. The authors used the Ibk algorithm to predict seven métiers from the daily boat record of landings in small-scale, multi-gear, multispecies fisheries. For each métier, a binary classification was completed, achieving an accuracy of 99% on a cross-validation procedure. Moreover, the authors directly apply the algorithm to non-validated data without performing a final test on unseen and validated data. To classify the non-validated trips, the authors first chose the most plausible method from the seven developed for predicting the métier, indicating that some human validation was still required to determine which algorithm to use. While the authors achieved good results, the absence of a dedicated test phase on unseen data suggests a less robust approach compared to the fully automated process implemented in the current work.

Moreover, Russo et al. (2010) implemented an Artificial Neural Network (ANN) to identify métiers using the gear licenses as variables [55]. Despite not having the same objective as the current thesis, the authors reached an accuracy of 94% over 15 métiers. In the current work, ANN were also implemented in a first analysis (in this case, Multi-Layer Perceptons), but did not achieve a similar performance (78%). Furthermore, Russo et al. (2016) implemented an innovative approach using Self-Organised Maps (SOM) to identify fishing gears using VMS and logbook data, which succeeded in correctly recognizing the gear in more than 82% of cases [54]. This study comprised nine gears of which, GTR, PS, and PTM illustrated to be less accurately recognized by the models. GTR were misclassified as LLS which rarely occurred in our analysis, which indicates how the fisheries may vary the patterns depending on the fishing areas. Nonetheless, PS and PTM were most of the time misclassified as LA, which is a gear that was not presented in the common framework between the Portuguese and Spanish datasets. Moreover, we may see these differences in the performance as the authors' data contains only Large-Scale Fisheries (LSF)(overall length > 12m).

Few works make use of landing profiles to identify fishing gears. Yet, several have been developed to identify fishing gears through tracking data, such as Vessel Monitoring Systems (VMS) or Automatic Identification Systems (AIS, which are only mandatory for vessels >= 10 meters). Kim et al (2020)[31], implemented Convolutional Neural Networks (CNN) to identify fishing gears using AIS data and reached an accuracy of 96.3% over 6 classes of fishing gears. In their work, the authors presented a higher misclassification between net classes and longlines, similar to what was stated by [54], reinforcing that LSF may differ in their behaviours compared to SSF. Furthermore, Marzuki et al. (2015)[38]

implemented Support Vector Machine (SVM) and Random Forest (RF) models reaching an accuracy of 94.59% over four fishing gears. The same authors later in 2017 [37], combined the previous models with mixture models, namely, GMM-VpVt, to identify the same fishing gears with an accuracy of 97.6% and 96.8% respectively. The authors further state that longlines and purse-seine were the gears that were more misclassified, which rarely occurs in the current work. Similar results were achieved by Carlos et al. (2022) (95% accuracy), which implemented several unsupervised and supervised algorithms such as K-Nearest Neighbours and SVMs [9]. Nevertheless, the current thesis achieved slightly better performances than the cited works, also permitting an extensive analysis of which species are most frequently caught in each fishery.

These latest works achieve good performances, allowing for a better spatial planning of the different fisheries, however, associating the fishing effort to the species caught would enhance that management and avoid stock depletion. The recent proposal that aims to amend Council Regulation (EC) No 1224/2009 (2023) includes mandatory tracking devices for Small-Scale Fisheries (SSF), and electronic logbook reporting, which would make the framework developed in the current thesis and the works cited above to be implemented together, strengthening the management of this fisheries that wield a substantial influence on the maritime ecosystem.

Explainable ML & AI - Feature Contribution

After implementing the models, it is crucial to understand the reasoning behind the algorithm's decision to classify an observation into a particular class. Nevertheless, recent studies have highlighted that this understanding becomes challenging when dealing with complex methodologies, often referred to as black-box models [34, 48]. Several interpretability tools, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), have been developed to provide insights into these models' decisions, simplifying the interpretation of complex algorithms.

SHAP provides consistent and theoretically sound explanations by considering all possible feature interactions, providing global and local interpretability, and making it versatile for various model types. On the other hand, LIME offers a more local perspective, generating explanations for individual predictions by approximating the model locally, which is ideal for understanding individual predictions in simpler models [15].

Alternatively, some practitioners opt for simpler, more interpretable models like Logistic Regression (LR) or suggest creating new interpretable models, even at the cost of potentially lower accuracy compared to more complex models like Random Forests, Artificial Neural Networks, or XGBoost [52].

In the current thesis, Shapley Additive Explanations (SHAP) were used to interpret the XGBoost model implemented as we want to understand which features are related

to the fishing gears both globally and individually. Additionally, a Logistic Regression model was trained on the same data to offer a more straightforward interpretation by analyzing the model's coefficients, which allows us to make direct inferences about the relationship between features and fishing gear classification.

Nonetheless, analysing SHAP and Logistic Regression (despite its lowest precision and recall) for each gear illustrated interesting conclusions.

Dredges

Using SHAP to understand the decision-making process of XGBoost illustrated that Dredges (DRB) had Chamelea and Donax as the two species that most contributed to a positive classification of the gear. These species are molluscs and are the most commonly caught within these fisheries [4, 13, 44]. Moreover, Logistic Regression was coherent with these studies and SHAP as Bolinus, Chamelea, and Donax were the species that contributed more to the classification of the gear.

These findings demonstrate that both models effectively capture the key characteristics of this fishery. By identifying Chamelea and Donax as significant contributors to gear classification, they align with known patterns of species catch in these fisheries, thus illustrating their ability to accurately reflect the fishery's dynamics.

Pots & Traps

The SHAP analysis for Pots and Traps (FPO) demonstrated that low octopus catches negatively impact the classification of this gear type. Additionally, the Logistic Regression (LR) model showed that octopus and sepia positively contribute to the classification. These findings align with previous studies conducted on the Algarve coast, which indicates that these fisheries typically catch Octopus and Sepia [47, 63, 65].

The SHAP analysis also suggests that these fisheries tend to have higher fishing activity in the first months of the year. However, Szynaka et al. (2021) reported that octopuses are consistently abundant throughout all seasons in the same coastal area [65]. This discrepancy implies that, according to the XGBoost model, octopuses are more likely to be caught earlier in the year compared to other species in our dataset. It's important to note that these insights might differ due to the use of data from similar, but not identical, fishing grounds, which could influence the analysis.

Overall, both SHAP and LR emphasize the findings of the cited studies, underscoring that FPO primarily catches octopuses.

Gillnets

The SHAP analysis identified Merluccius and Diplodus as the species that most positively contributed to the classification of Gillnets (GNS). However, the LR model only highlighted Merluccius as having a similar contribution. Merluccius has been frequently associated

with this type of fishery, suggesting that despite the lower accuracy in gear identification, both SHAP and LR correctly recognized that this species is commonly caught by fisheries using GNS [60].

The Food and Agriculture Organization of the United Nations (FAO) states that these fisheries mainly target demersal (refers to fish that live and feed near or on the seabed) or benthic (organisms that live on or near the bottom of a body of water) species, including Diplodus and Merluccius [1, 13], which supports the results from both SHAP and LR. Interestingly, although Sepia and Mullus are considered demersal species, their catch was associated with a negative contribution to the GNS classification. This suggests that, despite targeting demersal species, the fishery may be focusing on specific species.

Moreover, SHAP accurately identified a key characteristic of GNS fisheries—the tendency to catch multiple distinct species (taxa richness) within a single trip. This observation aligns with findings by Santos et al. (2002) and Dias et al. (2020), who noted that GNS fisheries typically capture several species [18, 60]. Additionally, the FAO mentions that this gear is capable of catching almost all species of fish and crustaceans, further illustrating the diversity inherent to GNS [13].

Trammel Nets

Taking into consideration the analysis of the SHAP for the trammel nets (GTR), it was illustrated that similarly to gillnets (GNS) this fishery tends to have a higher diversity of species caught within a trip. This was also a characteristic mentioned in some works such as Konstantinos et al. (2006) [64], Erzini et al. (2006) [21] and Batista et al. (2009)[6]. Another detail highlighted by SHAP, but not mentioned in these works, was that the overall length (LOA) of the fisheries using this gear tend to be smaller boats (within small-scale fisheries which already have an LOA < 12m). This was also evident in the data exploratory analysis in the previous Section 3.3, Figure 3.4.

Moreover, according to SHAP, Mullus and Sepia are the species that contribute the most to the classification of this gear. LR achieved similar conclusions, whereas Mullus, Sepia, and Pegusa were the species that most contributed to the classification of GTR. Similarly, Batista et al. (2009) state that in South Portugal these fisheries tend to catch more Sepia [6]. Erzini et al. (2006) also mention Sepia, Solea and Microchirus as the most common species caught by trammel nets in the Algarve [21].

Hand & Pole Lines

Relatively to Hand and Pole lines, to the author's best knowledge, there are no previous works on the characteristics of this fishery within the coastal area of Algarve or South Spain. Moreover, looking into detail to the FAO fishing gear descriptions and relation with species, LHP is one of the unique gears that does not have a detailed description about which species are caught [13].

Regardless, Punzón et al. (2004) analysed this fishery in Northwest Spain illustrating that these mainly catch Scomber [50]. These findings were not coherent with the SHAP analysis as this species worked as an indicator of not being a fishery using LHP. Furthermore, catching Pagellus and Dentex was a negative indicator too.

Moreover, the SHAP analysis highlighted a positive contribution of Mullus, Pagrus, Octopus, and Raja. In the Algarve, catching Octopus with LHP is not common within the Portuguese fleet, as the preferred method is Pots and Traps as illustrated previously. However, data from Spain illustrate that this fishery tends to catch a considerable amount of octopus with this gear. This may be due to this fishing gear being more selective allowing the fishermen to target only the Octopus species, which makes this technique more environmentally friendly compared to other fishing gear, consequently, it reduces bycatch (the capture of unintended species) [13]. Additionally, this pattern is verified by SHAP which indicates that having a greater taxa richness negatively impacts the model decision-making to classify a trip as LHP.

On the other hand, LR was completely out of the scope, highlighting that Chamelea had a positive contribution to the gear classification. This species is a bivalve which is not feasible to catch with this gear. This is explainable due to its performance on this gear (f1-score of 2%), and thereafter, these results are untrustful for LHP.

Longlines

The SHAP analysis of Longlines (LLS) illustrates that these fisheries tend to catch Pagrus, Pagellus, and Phycis. Additionally, Erzini et al. (1998, 2003, and 2010), analysed this fishery for several years, state that the most common species caught is Pagellus, which is coherent with the findings of this study [19, 20, 22].

Additionally, the taxa richness yields a negative contribution too. Logistic Regression has illustrated that Muraena and Phycis were the species that most contributed to a positive classification of LLS and that Lophius and Sepia would negatively influence the gear classification, which are typical benthic species.

Otter Bottom Trawllers

The SHAP analysis for Otter Bottom Trawllers (OTB) revealed that this gear tends to have larger boats operating it. This was also highlighted by the exploratory data analysis (EDA) in Section 3.3. Similarly, to the EDA, SHAP illustrated that this fishery catches several species, which have a small and positive contribution to the classification of the gear. Nonetheless, the species that positively indicate catches by this gear are Lophius, Octopus, Mullus, Trachinus, Sepia, and Microhirus. Moreover, Logistic Regression has illustrated a similar output, highlighting additional species such as Merluccius, Solea, and Loligo.

Furthermore, all these insights are coherent with other works such as Costa et al. (2008) who highlight that this fishery tends to catch Merluccius [16]. Campos et al. (2007) mention that a few of these fisheries tend to catch Octopuses and Sepia, which was noted by SHAP

[8]. These were not the species that contributed more, but they positively contributed to the decision-making process of the model towards this gear. Also, Alzorriz et al. (2016) [3] and Maynou et al. (2021) [39] state that boats using OTB as gear commonly catch Merluccius and Mullus.

All these works reinforce that both SHAP and Logistic Regression can capture the species and other variables that characterise OTB in south Spain and the Algarve.

Purse Seiners

SHAP analysis for Purse Seiners (PS) distinguishes this fishing gear by its' total weight in kilograms caught. This was also highlighted by Almeida et al. (2014) and Monteiro et al. (2016), who state that the species stocks caught by these fisheries are still exploited at levels that jeopardize the maximum sustainable yield [2, 43].

Moreover, SHAP highlights that catching Scomber, Trachurus, and Euthynnus positively contributes to identifying this gear. Logistic Regression provides a similar output highlighting additional species such as Mullus, Pomadasys, and Sardina. Despite LR performance on this gear, Mullus and Pomadasys were rarely caught by PS on the dataset, which may lead us to wrong conclusions. Nonethless, the other species mentioned were coherent with several studies performed on PS [2, 24, 43, 66].

Additionally, these fisheries tend to catch a lower variety of species as a higher taxa richness negatively impacts the classification of the gear. Nonetheless, catching Diplodus and Scorpaena indicates a negative contribution to identifying this gear.

Throughout this thesis, both Shapley Additive Explanations (SHAP) and Logistic Regression (LR) performed well in identifying the key variables that characterize fisheries, based on the models' respective performances and their alignment with previous studies discussed. Notably, SHAP provided a species-focused analysis while also highlighting other important factors such as taxa richness, the month of landing, and the boat's length. In contrast, LR offered insights that emphasized species-related variables more directly.

Many of the studies cited in this discussion chapter, primarily focus on specific fishing gear and target particular species, narrowing their scope to a more focused analysis. This is a common approach in fisheries research, where gear and species-specific studies often dominate. However, both SHAP and LR offer more flexibility, allowing for both targeted and general analyses. These methods enable a more comprehensive understanding of fisheries by incorporating broader factors like boat characteristics and seasonal trends in addition to species data, making them valuable tools for more holistic fisheries assessments.

Implications and Future Research

The current framework illustrates an excellent performance in identifying the fishing gears on the Spanish dataset and reliable performance when applied to the Portuguese

dataset (a data source with similar fishing grounds). These findings highlight a wide flexibility to apply this framework in other use cases such as the northwestern part of Spain and Portugal or even in institutions in other countries with similar fishing grounds. Furthermore, this framework could help validate trips within institutions that sometimes do not have enough resources to validate landed trips or, similarly to the use case in Portugal, when only the official licenses are provided without confirmation.

Relatively to the scalability of the models used, Random Forest (RF) and XGBoost (XGB), in the context of the current dataset, it's important to consider both the size and complexity of the data, as well as the computational resources required to process it. In this case, using a dataset of 29197 observations (75% of the observations - training data) with 75 variables (Species + Taxa Richness + LOA + TotalKg + Month of the Landing), the models ran in less than two minutes on a MacBook M1, but approximately one hour when performing the hyperparameter fine-tuning. This demonstrates that both models are efficient when handling moderate-sized datasets with high-dimensional data, even on consumer-grade hardware.

However, the scalability may become more challenging as the dataset grows in size, both in terms of the number of observations and features. While both RF and XGBoost can handle large datasets, XGBoost generally scales better due to its optimization techniques, such as tree pruning and parallel processing. On the other hand, Random Forest can become computationally expensive as the number of trees increases, particularly if the goal is to tune hyperparameters or explore different feature subsets.

For significantly larger datasets or higher-dimensional data, more robust infrastructure or cloud-based computing solutions may be required to maintain similar processing times. Additionally, model optimization techniques, such as feature reduction or sampling strategies, can help mitigate performance issues. Despite these considerations, the efficiency shown in this scenario indicates that for moderate datasets, RF and XGBoost can be applied without heavy computational demands, suggesting scalability to larger datasets with appropriate adjustments.

Beyond technical performance, another important aspect of this framework is its adaptability to specific fishing practices across regions. This raises important considerations about the species caught by different fisheries. Fishermen may choose types of gear based on their expertise to target the same species. For example, in Spain, both Pots & Traps and Long & Pole Lines are used to catch octopus, while in Portugal, Pots & Traps are the preferred method [4, 44]. Moreover, despite similar fishing grounds, some species are caught in one country but not in the other (see Annex I). Therefore, unifying the datasets is crucial to ensure that the framework is reliable and adaptable across different regions and use cases.

Moreover, the current model has limitations when it comes to capturing the full

diversity of Small-Scale Fisheries, especially in Portugal. These fisheries often hold licenses for multiple types of fishing gear, allowing them to use more than one type of gear on a single trip. However, the model currently only predicts the most likely gear used, which may not accurately reflect this practice. Future research should explore how often these fisheries use multiple gears in a single trip and determine whether it would be beneficial to develop a new multi-class framework to better capture this complexity.

Nonetheless, with the new amends to Council Regulation (EC) No 1224/2009 (2023) to have mandatory tracking devices and electronic reports of landing data within Small-Scale Fisheries (SSF) this framework could be easily applied. The high accuracy and adaptability of the model, demonstrate its potential to accurately classify fishing gears based on landing profiles and other variables. This level of precision can help regulate fishing efforts more effectively, minimizing overfishing by closely monitoring the use of gear types that target specific species. Furthermore, the insights into species catch, fishing gear, and the timing of operations provide a foundation for more sustainable resource allocation, ensuring that species with higher vulnerability to overfishing can be better protected.

Incorporating data from tracking devices into the framework could provide real-time monitoring of fishing activities. There are already highly effective methods for detecting fishing activity [7, 28, 29, 40, 53, 58, 41], by adding this information to the framework, authorities would be able to monitor compliance with fishing quotas and seasonal closures. Additionally, it would improve spatial management by creating a comprehensive framework that shows where boats are fishing, which species are being caught, and which gears are being used.

In the long term, this could lead to healthier fish stocks, greater biodiversity, and a more resilient ecosystem. It would also support the economic sustainability of small-scale fisheries. Ultimately, this approach offers a scalable solution that can be adapted to different regions and fishing gear types, making it a valuable tool for fisheries management worldwide.

7

Conclusions

Small-scale fisheries (SSF) wield substantial influence on the maritime ecosystem, however, until recent efforts to change Council Regulation (EC) No 1224/2009, they have been operating without obligations such as fishing authorisations, landing declarations, and sales notes. Without these restrictions, it is challenging to analyze logbooks or landing data, which was already difficult as the information registered would depend on the fishermen's credibility. Furthermore, in Portugal, the SSF usually operate with different gears, whereas only licenses are provided and not validated. Nonetheless, the combination of these factors makes it complex to understand the relationship between fishing gears and the target species.

In this thesis, to gain a better understanding of the Portuguese fleet behaviour through the data mentioned, in collaboration with the Instituto Español de Oceonografía (IEO) a validated dataset from South Spain was used to train supervised machine learning (ML) algorithms to identify the fishing gears, which were then applied to the Portuguese dataset.

The implementation of ML algorithms illustrated a good performance, namely, Random Forest (RF) and XGBoost (XGB) using the variables: weight in kilograms of the species caught (71 species), taxa richness, total catches in kg, the overall length of the boat and month of the landing. These algorithms could identify most of the fishing gears in south Spain with an accuracy of at least 94%. Nonetheless, Gillnets (GNS) were commonly mistaken as Trammel Nets (GTR), and after discussing the results with fisheries experts, it was concluded that these are very hard to distinguish by looking into the data provided. Hence, a new model was built whereas these gears were aggregated as one single class, achieving an accuracy of at least 99%. XGB performed slightly better overcoming the performance of RF only in recall by one percentage point.

Moreover, the trained XGB algorithm was used to identify the fishing gears in the Portuguese landing data, achieving a mean correspondence of 81% with the main licenses provided and a coherency of 83% with the fisheries experts' validation.

Afterwards, Shapley Additive Explanations (SHAP) and Logistic Regression (LR) were applied to understand the relationship between the variables and the fishing gears. In this

case, SHAP was applied to the results of XGB so we could have a better understanding of the model decision-making process. On the other hand, LR was applied to the landing data to provide a more direct and comprehensive approach to the interpretability of the relationship between the gears and the variables. The highlights of both models illustrated that Dredges were associated with catching Chamlea, Donax, and Bolinus. Pots & Traps were linked to Octopus, Sepia, and more activity early in the year. Gillnets were identified by catching Merluccius, Diplodus, and showing higher species richness. Trammel Nets stood out for higher species richness, catching Mullus, Sepia, Pegusa, and being used by smaller boats. Hand & Pole Lines were associated with catching Mullus, Pagrus, Octopus, and Raja, and lower species richness. Longlines were characterized by catching Pagrus, Pagellus, Phycis, and Muraena. Otter Bottom Trawlers were linked to larger boats, higher species richness, and catching Lophius, Octopus, Mullus, Trachurus, Sepia, Microchirus, Loligo, Solea, and Merluccius. Finally, Purse Seiners were characterized by a high total catch weight and species like Scomber, Trachurus, Euthynnus, Mullus, Pomadasys, and Sardina.

The current work demonstrates significant utility, achieving excellent performance on the original South Spain dataset and good performance on the Portuguese dataset, despite differences between them. This highlights the robustness of the trained models, even when applied to datasets from similar yet distinct fishing grounds. Additionally, the analysis provided valuable insights into the relationship between fishing gears and key variables, offering a deeper understanding of fleet behaviour and fishing practices that can help inform sustainable management decisions in small-scale fisheries.

Importantly, this work was recognized for its contribution to the field and was accepted and presented at the Annual Science Conference (ASC) 2024 of the International Council for the Exploration of the Sea (ICES), underscoring its relevance and impact within the scientific and fisheries management community [59]. Additionally, I participated in the ICES Workshop on Small Scale Fisheries and Geo-Spatial Data 2, hosted in Faro, in 2023, where I presented related-theme works, further contributing to the dialogue and research efforts in this area [29].

Bibliography

[1] S. Ahyong et al. *World Register of Marine Species (WoRMS)*. =https://www.marinespecies.org. Accessed: 2024-01-27. 2024-01-27. URL: https://www.marinespecies.org (cit. on pp. 8, 48).

[2] C. Almeida et al. "Environmental assessment of sardine (Sardina pilchardus) purse seine fishery in Portugal with LCA methodology including biological impact categories". In: *The International Journal of Life Cycle Assessment* 19 (2014), pp. 297–306. DOI: 10.1007/s11367-013-0646-5 (cit. on pp. 5, 50).

[3] N. Alzorriz et al. "Questioning the effectiveness of technical measures implemented by the Basque bottom otter trawl fleet: Implications under the EU landing obligation". In: *Fisheries Research* 175 (2016), pp. 116–126. DOI: 10.1016/J.FISHRES.2015.11.023 (cit. on p. 50).

[4] M. Anjos et al. "Bycatch and discard survival rate in a small-scale bivalve dredge fishery along the Algarve coast (southern Portugal)". In: *Scientia Marina* (2018). DOI: 10.3989/SCIMAR.04742.08A (cit. on pp. 5, 47, 51).

[5] M. Baki et al. "Fish species diversity, fishing gears and crafts from the Buriganga river, Dhaka". In: *Bangladesh Journal of Zoology* 45 (2017), pp. 11–26. DOI: 10.3329/BJZ.V45I1.34190 (cit. on p. 44).

[6] M. Batista, C. Teixeira, and H. Cabral. "Catches of target species and bycatches of an artisanal fishery: The case study of a trammel net fishery in the Portuguese coast". In: *Fisheries Research* 100 (2009), pp. 167–177. DOI: 10.1016/J.FISHRES.2009.07.007 (cit. on pp. 5, 48).

[7] F. Behivoke et al. "Estimating fishing effort in small-scale fisheries using GPS tracking data and random forests". In: *Ecological Indicators* 123 (2021), p. 107321. ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2020.107321. URL: https://www.sciencedirect.com/science/article/pii/S1470160X20312632 (cit. on p. 52).

[8] A. Campos et al. "Definition of fleet components in the Portuguese bottom trawl fishery". In: *Fisheries Research* 83 (2007), pp. 185–191. DOI: 10.1016/J.FISHRES.2006.09.012 (cit. on p. 50).

[9] H. Carlos et al. "Fishing Gear Pattern Recognition by Including Supervised Autoencoder Dimensional Reduction". In: *IEEE Geoscience and Remote Sensing Letters* 19 (2022), pp. 1–5. DOI: 10.1109/LGRS.2021.3084183 (cit. on pp. 5, 46).

[10] F. Chollet. *Deep Learning with Python*. 1st. USA: Manning Publications Co., 2017. ISBN: 1617294438 (cit. on p. 17).

[11] M. Choudhury, S. Paul, and D. B. Chhetri. "Study on fish catching devices used by the fishing community of Dewaddhar village of Sonebeel, Assam, India". In: *International Journal of Fisheries and Aquatic Studies* (2021). DOI: 10.22271/fish.2021.v9.i4d.2548 (cit. on p. 44).

[12] E. C. contributors. *Ocean and Fisheries, Sustainable fisheries, Rules, Small-scale fisheries*. 2024-01. URL: https://oceans-and-fisheries.ec.europa.eu/fisheries/rules/small-scale-fisheries_en (visited on 2024-01-26) (cit. on p. 1).

[13] E. C. contributors. *Oceans and Fisheries, Sustainable fisheries, Markets and trade, Seafood markets, Commercial designations, Fishing gears*. 2024-01. URL: https://fish-commercial-names.ec.europa.eu/fish-names/fishing-gears_en#GE (visited on 2024-01-26) (cit. on pp. 4, 8, 47–49).

[14] G. contributors. *Understanding Logistic Regression*. 2023-11. URL: https://www.geeksforgeeks.org/understanding-logistic-regression/ (visited on 2024-06-05) (cit. on p. 26).

[15] markovML contributors. *LIME vs SHAP: A Comparative Analysis of Interpretability Tools*. 2024-02. URL: https://www.markovml.com/blog/lime-vs-shap (visited on 2024-08-23) (cit. on p. 46).

[16] M. E. Costa, K. Erzini, and T. Borges. "Bycatch of crustacean and fish bottom trawl fisheries from southern Portugal (Algarve)". In: *Scientia Marina* 72 (2008), pp. 801–814. DOI: 10.3989/SCIMAR.2008.72N4801 (cit. on pp. 5, 49).

[17] N. Das et al. "Explaining predictions of an automated pulmonary function test interpretation algorithm". In: *M-health/e-health* (2019). DOI: 10.1183/13993003.congress-2019.pa2227 (cit. on p. 5).

[18] V. Dias et al. "High Coral Bycatch in Bottom-Set Gillnet Coastal Fisheries Reveals Rich Coral Habitats in Southern Portugal". In: *Frontiers in Marine Science* 7 (2020). DOI: 10.3389/fmars.2020.603438 (cit. on p. 48).

[19] K. Erzini et al. "Competition between static gear of the small-scale fisheries in Algarve waters (southern Portugal)". In: *Mediterranean Marine Science* 11 (2010), pp. 225–244. DOI: 10.12681/MMS.74 (cit. on pp. 5, 49).

[20] K. Erzini et al. "Quantifying the roles of competing static gears: comparative selectivity of longlines and monofilament gill nets in a multi-species fishery of the Algarve (southern Portugal)". In: *Scientia Marina* 67 (2003), pp. 341–352. DOI: 10.3989/SCIMAR.2003.67N3341 (cit. on p. 49).

[21] K. Erzini et al. "Size selectivity of trammel nets in southern European small-scale fisheries". In: *Fisheries Research* 79 (2006), pp. 183–201. DOI: 10.1016/J.FISHRES.2006.03.004 (cit. on p. 48).

[22] K. Erzini et al. "Species and size selectivity in a 'red' sea bream longline 'métier' in the Algarve (southern Portugal)". In: *Aquatic Living Resources* 11 (1998), pp. 1–11. DOI: 10.1016/S0990-7440(99)80025-4 (cit. on p. 49).

[23] M. FAO. "Strategies for increasing the sustainable contribution of small-scale fisheries to food security and poverty alleviation". In: *Report of the twenty-fifth session of the Committee on Fisheries* (2003), pp. 76–84 (cit. on p. 1).

[24] D. Feijó et al. "Trends in the activity pattern, fishing yields, catch and landing composition between 2009 and 2013 from onboard observations in the Portuguese purse seine fleet". In: *Regional Studies in Marine Science* (2018). DOI: 10.1016/J.RSMA.2017.12.007 (cit. on p. 50).

[25] J. Friedman. "Greedy function approximation: A gradient boosting machine." In: *Annals of Statistics* 29 (2001), pp. 1189–1232. DOI: 10.1214/AOS/1013203451 (cit. on p. 17).

[26] A. Geron. *Hands-on machine learning with scikit-learn, keras, and TensorFlow 3e: Concepts, tools, and techniques to build intelligent systems*. en. 3rd ed. Sebastopol, CA: O'Reilly Media, 2022. ISBN: 9781098125974 (cit. on pp. 14, 15, 17, 19, 20).

[27] H. Hakkoum, I. Abnane, and A. Idri. "Evaluating Interpretability of Multilayer Perceptron and Support Vector Machines for Breast Cancer Classification". In: *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)* (2022), pp. 1–6. DOI: 10.1109/AICCSA56895.2022.10017521 (cit. on p. 5).

[28] ICES. "Workshop on Geo-Spatial Data for Small-Scale Fisheries (WKSSFGEO)". In: (2022-02). DOI: 10.17895/ices.pub.10032. URL: https://ices-library.figshare.com/articles/report/Workshop_on_Geo-Spatial_Data_for_Small-Scale_Fisheries_WKSSFGEO_/19248947 (cit. on p. 52).

[29] ICES. "Workshop on Small Scale Fisheries and Geo-Spatial Data 2 (WKSSFGEO2)". In: (2024-01). DOI: 10.17895/ices.pub.22789475.v1. URL: https://ices-library.figshare.com/articles/report/Workshop_on_Small_Scale_Fisheries_and_Geo-Spatial_Data_2_WKSSFGEO2_/22789475 (cit. on pp. 52, 54).

[30] S. ben Jabeur, S. Mefteh-Wali, and J. Viviani. "Forecasting gold price with the XGBoost algorithm and SHAP interaction values". In: *Annals of Operations Research* (2021). DOI: 10.1007/S10479-021-04187-W (cit. on p. 5).

[31] K.-i. Kim and K. M. Lee. "Convolutional Neural Network-Based Gear Type Identification from Automatic Identification System Trajectory Data". In: *Applied Sciences* 10.11 (2020). ISSN: 2076-3417. DOI: 10.3390/app10114010. URL: https://www.mdpi.com/2076-3417/10/11/4010 (cit. on pp. 4, 45).

[32] P. Leitão et al. "Time and spatial trends in landing per unit of effort as support to fisheries management in a multi-gear coastal fishery". In: *PLOS ONE* 17.7 (2022-07), pp. 1–20. DOI: 10.1371/journal.pone.0258630. URL: https://doi.org/10.1371/journal.pone.0258630 (cit. on pp. 1, 8, 10, 43).

[33] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

[34] O. Loyola-González. "Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View". In: *IEEE Access* 7 (2019), pp. 154096–154113. DOI: 10.1109/ACCESS.2019.2949286 (cit. on p. 46).

[35] S. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: 2017-12 (cit. on p. 23).

[36] P. Marchal. "A comparative analysis of métiers and catch profiles for some French demersal and pelagic fleets". In: *ICES Journal of Marine Science* 65.4 (2008-03), pp. 674–686. ISSN: 1054-3139. DOI: 10.1093/icesjms/fsn044. eprint: https://academic.oup.com/icesjms/article-pdf/65/4/674/29130834/fsn044.pdf. URL: https://doi.org/10.1093/icesjms/fsn044 (cit. on p. 1).

[37] M. I. Marzuki et al. "Fishing Gear Identification From Vessel-Monitoring-System-Based Fishing Vessel Trajectories". In: *IEEE Journal of Oceanic Engineering* 43.3 (2018), pp. 689–699. DOI: 10.1109/JOE.2017.2723278 (cit. on pp. 4, 46).

[38] M. I. Marzuki et al. "Fishing gear recognition from VMS data to identify illegal fishing activities in Indonesia". In: *OCEANS 2015 - Genova*. 2015, pp. 1–5. DOI: 10.1109/OCEANS-Genova.2015.7271551 (cit. on pp. 4, 45).

[39] F. Maynou et al. "Relative Catch Performance of Two Gear Modifications Used to Reduce Bycatch of Undersized Fish and Shrimp in Mediterranean Bottom Trawl Fisheries". In: *Marine and Coastal Fisheries* (2021). DOI: 10.1002/mcf2.10178 (cit. on p. 50).

[40] T. Mendo et al. "Estimating fishing effort from highly resolved geospatial data: Focusing on passive gears". In: *Ecological Indicators* 154 (2023), p. 110822. ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2023.110822. URL: https://www.sciencedirect.com/science/article/pii/S1470160X23009640 (cit. on p. 52).

[41] T. Mendo et al. "Effect of temporal and spatial resolution on identification of fishing activities in small-scale fisheries using pots and traps". In: *ICES Journal of Marine Science* 76.6 (2019-04), pp. 1601–1609. ISSN: 1054-3139. DOI: 10.1093/icesjms/fsz073. eprint: https://academic.oup.com/icesjms/article-pdf/76/6/1601/31247341/fsz073.pdf. URL: https://doi.org/10.1093/icesjms/fsz073 (cit. on p. 52).

[42] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022. URL: https://christophm.github.io/interpretable-ml-book (cit. on p. 23).

[43] P. Monteiro. "The Purse Seine Fishing of Sardine in Portuguese Waters: A Difficult Compromise Between Fish Stock Sustainability and Fishing Effort". In: *Reviews in Fisheries Science Aquaculture* 25 (2017), pp. 218–229. DOI: 10.1080/23308249.2016.1269720 (cit. on p. 50).

[44] L. Nicolau et al. "Hand dredging for the wedge clam (Donax trunculus) in the Algarve coast (southern Portugal): fishing yield, bycatch, discards and damage rates". In: *Marine Biology Research* 17 (2021), pp. 960–977. DOI: 10.1080/17451000.2022.2048670 (cit. on pp. 47, 51).

[45] M. Palmer et al. "Combining sale records of landings and fishers knowledge for predicting métiers in a small-scale, multi-gear, multispecies fishery". In: *Fisheries Research* 195 (2017), pp. 59–70. ISSN: 0165-7836. DOI: https://doi.org/10.1016/j.fishres.2017.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0165783617301753 (cit. on pp. 3, 45).

[46] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 20).

[47] F. Pereira et al. "Catches of Sepia officinalis in the small-scale cuttlefish trap fishery off the Algarve coast (southern Portugal)". In: *Fisheries Research* (2019). DOI: 10.1016/J.FISHRES.2019.01.022 (cit. on pp. 5, 47).

[48] J. Petch, S. Di, and W. Nelson. "Opening the black box: the promise and limitations of explainable machine learning in cardiology." In: *The Canadian journal of cardiology* (2021). DOI: 10.1016/j.cjca.2021.09.004 (cit. on p. 46).

[49] C. Pita and M. Gaspar. *Small-Scale Fisheries in Portugal: Current Situation, Challenges and Opportunities for the Future*. Springer, 2020, pp. 283–305. DOI: 10.1007/978-3-030-37371-9_14 (cit. on p. 1).

[50] A. Punzón, B. Villamor, and I. Preciado. "Analysis of the handline fishery targeting mackerel (Scomber scombrus, L.) in the North of Spain (ICES Division VIIIbc)". In: *Fisheries Research* 69 (2004), pp. 189–204. DOI: 10.1016/J.FISHRES.2004.05.002 (cit. on pp. 5, 49).

[51] J. M. Rodriguez-Albala, A. Pea, P. Melzi, et al. "Spatio-temporal trajectory data modeling for fishing gear classification". In: *Pattern Anal Applic* 27 42 (2024). DOI: https://doi.org/10.1007/s10044-024-01263-2 (cit. on p. 4).

[52] C. Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1 (2018), pp. 206–215. DOI: 10.1038/s42256-019-0048-x (cit. on p. 46).

[53] M. M. Rufino et al. "Estimating fishing effort in small-scale fisheries using high-resolution spatio-temporal tracking data (an implementation framework illustrated with case studies from Portugal)". In: *Ecological Indicators* 154 (2023), p. 110628. ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2023.110628. URL: https://www.sciencedirect.com/science/article/pii/S1470160X23007707 (cit. on pp. 1, 52).

[54] T. Russo et al. "Modeling landings profiles of fishing vessels: An application of Self-Organizing Maps to VMS and logbook data". In: *Fisheries Research* 181 (2016), pp. 34–47. ISSN: 0165-7836. DOI: https://doi.org/10.1016/j.fishres.2016.04.005. URL: https://www.sciencedirect.com/science/article/pii/S016578361630100X (cit. on pp. 4, 8, 45).

[55] T. Russo et al. "When behaviour reveals activity: Assigning fishing effort to métiers based on VMS data using artificial neural networks". In: *Fisheries Research* 111.1 (2011), pp. 53–64. ISSN: 0165-7836. DOI: https://doi.org/10.1016/j.fishres.2011.06.011. URL: https://www.sciencedirect.com/science/article/pii/S0165783611002281 (cit. on pp. 4, 45).

[56] C. Sağlam and O. Akyol. "Diversity of used fishing gears in the Aegean lagoons". In: *Ege Journal of Fisheries and Aquatic Sciences* (2022). DOI: 10.12714/egejfas.39.1.02 (cit. on p. 44).

[57] S. Salas, E. Torres-Irineo, and E. Coronado. "Towards a métier-based assessment and management approach for mixed fisheries in Southeastern Mexico". In: *Marine Policy* 103 (2019), pp. 148–159. ISSN: 0308-597X. DOI: https://doi.org/10.1016/j.marpol.2019.02.040. URL: https://www.sciencedirect.com/science/article/pii/S0308597X18302999 (cit. on p. 1).

[58] N. Sales Henriques et al. "An approach to map and quantify the fishing effort of polyvalent passive gear fishing fleets using geospatial data". In: *ICES Journal of Marine Science* 80.6 (2023-06), pp. 1658–1669. ISSN: 1054-3139. DOI: 10.1093/icesjms/fsad092. eprint: https://academic.oup.com/icesjms/article-pdf/80/6/1658/51096632/fsad092.pdf. URL: https://doi.org/10.1093/icesjms/fsad092 (cit. on p. 52).

[59] J. Samarão. "Unveiling the gears in Small-scale fisheries using landings from a different country: an ML approach". In: *Proceedings of the Annual Science Conference (ASC) 2024, International Council for the Exploration of the Sea (ICES) - Theme Session Q: Small Scale-Fisheries where are you?* Presented at the Annual Science Conference (ASC) 2024. Gateshead, UK, 2024-09 (cit. on p. 54).

[60] M. N. Santos et al. "Gill net and long-line catch comparisons in a hake fishery: the case of southern Portugal". In: *Scientia Marina* 66 (2002), pp. 433–441. DOI: 10.3989/SCIMAR.2002.66N4433 (cit. on pp. 5, 48).

[61] L. S. Shapley. "17. A Value for n-Person Games". In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by H. W. Kuhn and A. W. Tucker. Princeton: Princeton University Press, 1953, pp. 307–318. ISBN: 9781400881970. DOI: doi:10.1515/9781400881970-018. URL: https://doi.org/10.1515/9781400881970-018 (cit. on p. 23).

[62] G. Shester and F. Micheli. "Conservation challenges for small-scale fisheries: Bycatch and habitat impacts of traps and gillnets". In: *Biological Conservation* 144 (2011), pp. 1673–1681. DOI: 10.1016/J.BIOCON.2011.02.023 (cit. on p. 44).

[63] C. P. Sonderblohm et al. "Participatory assessment of management measures for octopus vulgaris pot and trap fishery from southern Portugal". In: *Marine Policy* 75 (2017), pp. 133–142. DOI: 10.1016/J.MARPOL.2016.11.004 (cit. on p. 47).

[64] K. I. Stergiou et al. "Trammel net catch species composition, catch rates and métiers in southern European waters: A multivariate approach". In: *Fisheries Research* 79 (2006), pp. 170–182. URL: https://api.semanticscholar.org/CorpusID:83523882 (cit. on p. 48).

[65] M. J. Szynaka et al. "Identifying Métiers Using Landings Profiles: An Octopus-Driven Multi-Gear Coastal Fleet". In: *Journal of Marine Science and Engineering* 9.9 (2021). ISSN: 2077-1312. DOI: 10.3390/jmse9091022. URL: https://www.mdpi.com/2077-1312/9/9/1022 (cit. on pp. 2, 3, 10, 44, 47).

[66] R. Tejerina et al. "The purse-seine fishery for small pelagic fishes off the Madeira Archipelago". In: *African Journal of Marine Science* 41 (2019), pp. 373–383. DOI: 10.2989/1814232X.2019.1678520 (cit. on p. 50).

[67] C. Yang, M. Chen, and Q. Yuan. "The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis." In: *Accident; analysis and prevention* 158 (2021), p. 106153. DOI: 10.1016/j.aap.2021.106153 (cit. on p. 5).

[68] A. Zhang et al. *Dive into Deep Learning*. https://D2L.ai. Cambridge University Press, 2023 (cit. on pp. 19, 20).

[69] X.-q. Zhu et al. "An interpretable stacking ensemble learning framework based on multi-dimensional data for real-time prediction of drug concentration: The example of olanzapine". In: *Frontiers in Pharmacology* 13 (2022). DOI: 10.3389/fphar.2022.975855 (cit. on p. 5).

I

Annex 1

Table I.1: Comparison of species caught in Portuguese and South Spanish small-scale fisheries, highlighting species unique to each region and those common to both

Species caught Spain but not in Portugal			Species caught in Portugal but not in Spain			Species caught in common between Spain and Portugal					
Acanthocardia	Dactylopterus	Pasiphaea	Alosa	Exocoetidae	Platichthys	Argyrosomus	Dicentrarchus	Lithognathus	Octopus	Sarpa	Torpedo
Alloteuthis	Engraulis	Penaeus	Anguilla	Gaidropsarus	Pleuronectes	Auxis	Dicologlossa	Loligo	Pagellus	Scomber	Trachinus
Arnoglossus	Eutrigla	Peristedion	Anthias	Gymnura	Pseudotolithus	Balistes	Diplodus	Lophius	Pagrus	Scophthalmus	Trachurus
Bothus	Geryon	Plesionika	Beryx	Isurus	Pseudupeneus	Belone	Donax	Maja	Palinurus	Scorpaena	Umbrina
Calappa	Gobius	Sardinella	Bothidae	Menidia	Sebastes	Bolinus	Epinephelus	Merluccius	Pegusa	Scyliorhinus	Uranoscopus
Callista	Illex	Scyllarides	Callinectes	Muraenidae	Spisula	Boops	Euthynnus	Microchirus	Phycis	Sepia	Zeus
Carcinus	Labrus	Squilla	Centrolabrus	Murex	Stromateus	Brama	Galeus	Mugil	Plectorhinchus	Seriola	
Centrolophus	Lichia	Symphodus	Charonia	Necora	Trichiurus	Caranx	Halobatrachus	Mullus	Polyprion	Serranus	
Centrophorus	Liocarcinus	Thunnus	Cynoglossus	Ommastrephes	Trisopterus	Chamelea	Helicolenus	Muraena	Pomadasys	Solea	
Cepola	Osteichthyes	Todarodes	Cynoscion	Orcynopsis	Zenopsis	Chelidonichthys	Homarus	Mustelus	Pomatomus	Sparus	
Citharus	Parapandalus	Todaropsis	Dasyatis	Oxynotus	Zenopsis	Chelon	Lepidorhombus	Myliobatis	Prionace	Sphyraena	
Coryphaena	Parapenaeus	Trachinotus	Dipturus	Palinuridae		Conger	Lepidotrigla	Nephrops	Raja	Spicara	
Trigla	Venus	Xyrichtys	Epigonus	Parapristipoma		Dentex	Leucoraja	Oblada	Sardina	Spondyliosoma	

II

Annex 2

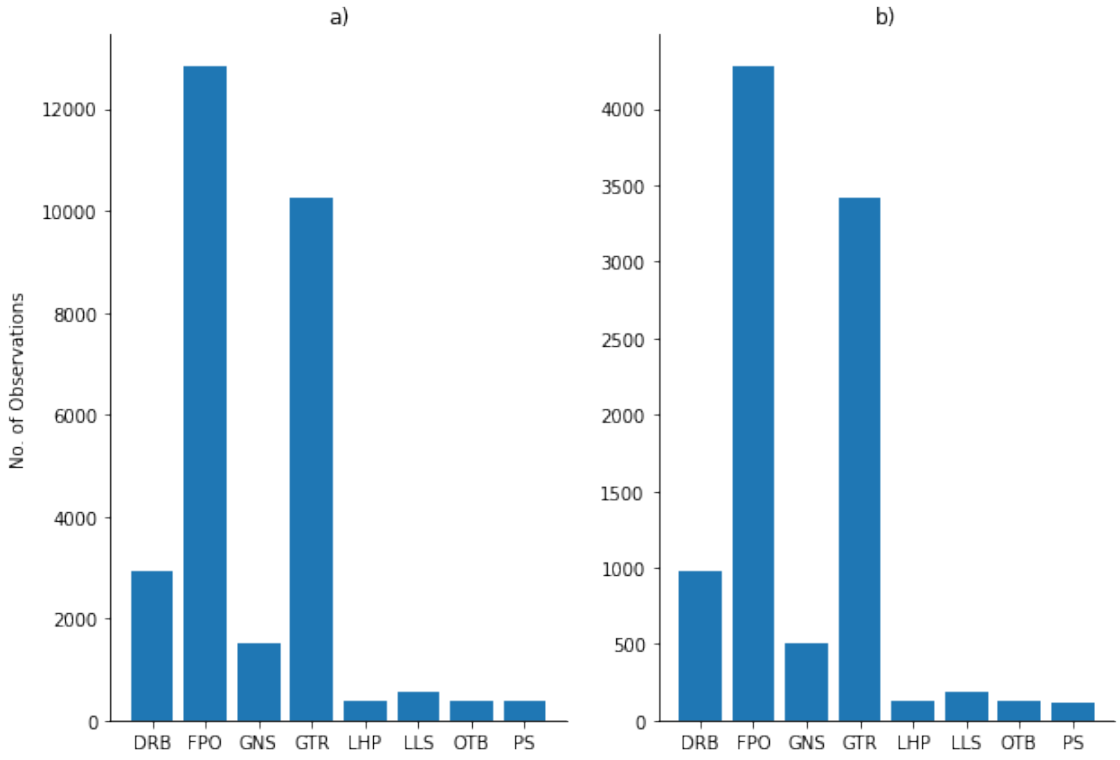


Figure II.1: Number of observations that were split into a) training and b) testing

III

Annex 3

Table III.1: Multi-layer Perceptons (MLP) performance, when using the taxa caught in kg, taxa richness (S), LOA, total kg caught by trip (independently of the taxa) and month of the landing to identify gears within IEO landing data.

Gear / Model	MLP F1-Score
DRB	68%
FPO	87%
GNS	9%
GTR	83%
LHP	9%
LLS	23%
OTB	79%
PS	69%
Accuracy	78%
Avg. Precision	69%
Avg. Recall	52%