Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Engineering
Department of Marine Technology

Beatrice Kiær

Prediction of Vessel Activity in Arctic Waters using Supervised Learning Methods

Master's thesis in Engineering and ICT
Supervisor: Ingrid Bouwer Utne
Co-supervisor: Alun Jones
June 2022

NTNU
Norwegian University of
Science and Technology

Beatrice Kiær

Prediction of Vessel Activity in Arctic Waters using Supervised Learning Methods

Master's thesis in Engineering and ICT
Supervisor: Ingrid Bouwer Utne
Co-supervisor: Alun Jones
June 2022

Norwegian University of Science and Technology
Faculty of Engineering
Department of Marine Technology

NTNU
Norwegian University of
Science and Technology

Preface

This master's thesis finalizes my studies at the MSc program in Engineering and ICT at NTNU with a chosen main profile within Marine Technology. The report is written during the spring semester of 2022 under the Department of Marine Technology and builds on findings from the specialization project carried out in the previous semester.

I would like to express my sincere thanks to my supervisors Alun Jones and Ingrid Bouwer Utne, for motivation, insightful discussions, and guidance related to the project scope, data sources, and implementations. I would also like to thank Benjamin Lagemann for providing access to the necessary hardware and always being available for technical assistance. Final thanks are given to my family and fellow students for continuous motivational support throughout the semester.

This thesis investigates the use of machine learning for predicting vessel activity in Arctic waters, represented by a selected and specified area in the Barents Sea. The activity measures are based on the authorized use of data obtained from the Arctic Council Working Group on the Protection of the Arctic Marine Environment's Arctic Ship Traffic Data (ASTD) System. Information on the ASTD System is available at http://www.astd.is. The relevant code implementations are attached in a separate compressed file in the submission.

Beatrice Kiær
June 18, 2022
NTNU

Abstract

The Arctic is undergoing rapid environmental changes by warmer temperatures and vanishing sea ice, leading to longer navigation seasons and alternative transportation routes. Increased maritime traffic in the harsh climate conditions leads to a higher probability of incidents that threaten people and the environment. A thorough understanding of the evolving activity is necessary to mitigate these threats. However, such insight is difficult to obtain from limited operational experiences as the region has historically been inaccessible. Meanwhile, today's technology offers new opportunities through software and high-quality data from sensors and satellites. The objective of this thesis is to investigate how machine learning (ML) and real-time vessel traffic data can be used for identifying activity trends in the Arctic and how these trends are connected with the Arctic climate changes, environment, and ecology.

The proposed solution provides the necessary steps from data allocation and processing to ML model development. In this thesis, the Arctic is represented by a selected area of the Barents Sea. The relevant period is between 2015 and 2021. Numerous data sources were investigated for potential ML model predictors. Data about temperature, sea ice, ocean depth, distance to land, and fish catches were found to be the most applicable according to file size, access, and coverage. The data were individually analyzed and modified, and the final ML training data set was composed through aggregation by a created spatiotemporal grid. The prediction target of activity was extracted from vessel traffic data, from which two case studies of different activity viewpoints were defined. The training data of each case study were used to build a Random Forest (RF) classifier and an Extreme Gradient Boosting (XGBoost) classifier. Whereas the first case study investigated the prediction of vessel presence, the latter predicted the level of vessel density by operational industry.

The prediction results show that the classifiers behaved similarly during learning and provided satisfactory performance in precision and recall, mostly above 80% in both case studies. This verifies the applicability of ML in terms of activity estimation. The detection of vessel presence achieved the best overall evaluation with an F1-score of 87%. Independently of target prediction and industry, *latitude* had the highest impact on prediction outcome. However, in contrast to expectations from the literature, neither a remarkable correlation between the activity density and climatic features nor an increase in activity over time was observed. This is most likely related to a short time frame investigated and a high concentration of vessel presence at the lower latitudes, where climatic changes are more stable than in the north. The learning process indicated predictions under uncertainty when the spatial separation between the target classes was less clear. Alternative regions and time frames should be analyzed, and further investigation of activity-influencing factors is needed to increase model confidence and capture the trends as expected.

Sammendrag

Raske miljøendringer i Arktis, i form av varmere temperaturer og smelting av havis, fører til lengre navigasjonssesonger og alternative transportveier. Økt sjøtrafikk i de tøffe klimaforholdene fører til en høyere sannsynlighet for at ulykker inntreffer, hvilket truer mennesker og miljø. En grundig forståelse av aktivitetsutviklingen er nødvendig for å redusere disse truslene. Slik innsikt er imidlertid vanskelig å oppnå gjennom tidligere erfaringer da Arktis opprinnelig har vært utilgjengelig. Samtidig tilbyr dagens teknologi nye muligheter gjennom programvare og høykvalitetsdata fra sensorer og satellitter. Formålet med denne masteroppgaven er å undersøke hvordan maskinlæring (ML) og trafikkdata fra fartøy i sanntid kan brukes til å identifisere aktivitetstrender i Arktis og hvordan disse trendene henger sammen med arktiske klimaendringer, miljø og økologi.

Metoden som presenteres følger stegene i en typisk ML-sekvens fra allokering og prosessering av data til implementering av to ML-modeller. I denne oppgaven er Arktis representert som et utvalgt område i Barentshavet. Den aktuelle perioden er mellom 2015 og 2021. Flere datakilder ble undersøkt for potensielle inngangsverdier til ML modellene, hvorav data om havtemperatur, havis, dybde, avstand til land og fiskefangst var de mest anvendelige i forhold til filstørrelse, tilgang og utstrekning. Dataene ble individuelt analysert og bearbeidet, og det endelige treningsdatasettet ble komponert gjennom aggregering til et forhåndsdefinert rutenett i tid og rom. Utgangsverdien for aktivitet ble definert fra trafikkdata gjennom to ulike case-studier. Treningsdataene fra hver case-studie ble brukt til å bygge en Random Forest (RF) modell og en Extreme Gradient Boosting (XGBoost) modell. Den første case-studien undersøkte predikering av aktivitet gjennom tilstedeværelse av fartøy, og den andre estimerte tettheten av fartøy innenfor en bestemt type skipsaktivitet.

Resultatene viser at modellene hadde tilnærmet lik læringsprosess og predikerte med tilfredsstillende verdier av presisjon og dekning, i snitt på over 80% i begge case-studiene. Dette illustreter potensialet av å bruke ML for estimering av aktivitet med tilsvarende inngangsverdier. Prediksjon av tilstedeværelse oppnådde den beste samlede evalueringen med en F1-score på 87%. Uavhengig av utgangsverdi og type skipsaktivitet, hadde *breddegrad* størst innvirkning på prediksjonsresultatene. I motsetning til forventninger fra litteraturen ble det imidlertid verken observert en tydelig korrelasjon mellom aktivitetstetthet og klimatiske forhold, eller en økning i aktivitet over tid. Dette er mest sannsynlig knyttet til den korte tidsrammen undersøkt og høy tilstedeværelse av fartøy ved lavere breddegrader, hvor klimaendringene er mer stabile enn i nord. Læringsprosessen indikerte prediksjoner under usikkerhet når det romlige skillet mellom utgangsverdiene var mindre tydelig. Alternative områder og tidsrammer bør analyseres, og ytterligere undersøkelser av påvirkningsfaktorer for aktivitet er nødvendig for å øke modellens pålitelighet og fange opp trendene som forventet.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
AIS	Automatic Identification System
AMSA	Arctic Marine Shipping Assessment
ANN	Artificial Neural Network
API	Application Programming Interface
ASTD	Arctic Ship Traffic Data
CSV	Comma Separated Value
EDA	Exploratory Data Analysis
EPPR	Emergency Prevention, Preparedness and Response
EU	European Union
FTP	File Transfer Protocol
GB	Gigabyte
IBCAO	International Bathymetric Chart of the Arctic Ocean
IMO	International Maritime Organization
ITU	International Telecommunication Union
MARPOL	International Convention of the Prevention of Pollution from Ships
MDI	Mean Decrease in Impurity
ML	Machine Learning
MMSI	Maritime Mobile Service Identity
NaN	Not a Number
NASA	National Aeronautics and Space Administration
NCA	Norwegian Coastal Administration
NetCDF	Network Common Data Form
NMDC	Norwegian Marine Data Centre
NSIDC	National Snow and Ice Data Center
NSR	Northern Sea route
NWP	Northwest Passage
OBPG	Ocean Biology Processing Group
PAME	Protection of the Arctic Marine Environment
RF	Random Forest
RIF	Risk Influencing Factor
SAR	Search and Rescue
SIC	Sea ice concentrations
SMOTE	Synthetic Minority Oversampling Technique
SOLAS	International Convention for the Safety of Life At Sea
SST	Sea surface temperatures
SVM	Support Vector Machine
TSR	Transpolar Sea Route
UNCLOS	United Nations Convention on the Law of the Sea
XGBoost	Extreme Gradient Boosting

Chapter 1

Introduction

This thesis aims to investigate the use of machine learning (ML) and traffic data recorded from vessels for identifying activity trends in the Arctic waters and how activity is related to the Arctic surroundings and climate change. The following chapter provides a motivational introduction and background of the thesis' objective, followed by a description of the work's scope and limitations and a brief report outline.

1.1 Background and motivation

Since 1979 around 3.49 million km^2 of the Arctic sea ice has melted. The resulting bare oceans absorb the sun's energy instead of reflecting it back, inducing a faster temperature increase and melting of sea ice, permafrost, and snow cover. As the Arctic plays a crucial role in climate regulations through the Earth's oceanographic and atmospheric circulations, the melting yields numerous impacts on weather, sea level, species, and habitats around the world (Ocean Conservancy 2017). At the same time, warmer seas lengthen the navigation season and enable alternative sea routes linking the Atlantic and Pacific oceans. Such increased access and discoveries of oil and gas resources lead to employment opportunities and facilitate financial and time savings (Marsh Risk Management Research 2014). As a result, maritime transportation and industrial operations are encouraged to enter the region, from which a response of increasing activity has been analyzed and confirmed by several studies (Azzara, H. Wang, and Rutherford 2019; Paxian et al. 2010; Stephenson, Smith, and Brigham 2013; M. Liu and Kronbak 2010).

The Arctic is characterized by storms, rough seas, variable sea ice, low temperatures, remoteness, and poor visibility. Either individually or combined, these characteristics have

the potential to damage humans and the environment with severe consequences according to the region's vulnerability. Hence, the Arctic involves a complex risk picture, which causes challenges concerning maritime infrastructure, emergency preparedness, and search and rescue (SAR) strategies (EPPR 2018). As a result, studies of the changing access to the Arctic focusing on operational preparedness and risk assessment have received more attention during the last two decades (Marchenko et al. 2018; Marchenko 2019; Benz, Münch, and Hartmann 2021). One example of such a study published by the Arctic Council in 2009 resulted in the first Arctic Marine Shipping Assessment (AMSA) report, which addresses Arctic shipping, maritime infrastructure, and associated environmental and human impacts (Arctic Council 2009). Since then, organizations and governments have been involved in the Arctic Council's work regarding Arctic risk-reducing measures. The AMSA report was a catalyst toward global regulations responding to the increased Arctic shipping, such as the mandatory Polar Code framework for vessel operations in Polar regions (IMO 2017). Implementations of such regulations and increased international collaboration form a critical step toward safe ship operations and environmental protection within the Arctic waters. However, today's research and risk assessment approaches rely heavily on past experiences and incidents, making it challenging to address all Arctic-related safety concerns and environmental threats due to the region's historical inaccessibility and lack of benchmarked data (Rusten et al. 2015).

Because of the weak knowledge base, concerns related to the Arctic should be addressed by alternative approaches that are not necessarily dependent on historical experiences. According to theory, *risk* can be expressed as a function of an incident's frequency multiplied by its associated consequences (IMO 2007). Implicitly, this means that activity is a prerequisite for the presence of risk. The more vessels are operating in the Arctic region, the higher the likelihood of incident occurrences that may affect human lives, the vulnerable Arctic environment, and ecosystems (Peters et al. 2011; Ocean Conservancy 2017). One possible strategy for minimizing these threats and improving knowledge of Arctic-associated risks is to target the Arctic traffic volumes in a broad context by investigating the individual sources of traffic growth. Such sources include both impacts from climate change, such as warmer temperatures and diminishing sea ice, and impacts from commercial operations related to the creation of new infrastructure and discoveries of gas and oil sources.

Today's technology provides enormous volumes of sensor data, including real-time marine traffic data from Automatic Identification Systems (AIS) and satellite data deriving information about the Earth's surface and environment. Moreover, with the evolving technology trends, there is an increased interest in how modern methodologies and improved data access may

facilitate and streamline industry procedures. ML is one such computer-based methodology for data analysis, which automatically learns patterns from large amounts of data and applies these patterns to new data for future problem solving (Russel and Norvig 2020a). The benefit of ML is its potential to find complex patterns without making assumptions that otherwise is challenging by using traditional statistical methods. While learning algorithms are increasingly being used to facilitate risk assessment within, for example, the financial, medical, and automotive industries, the use of ML within the maritime domain has received less attention. However, the increasing amounts of real-time traffic data from AIS carried by vessels combined with data sources describing the Arctic climate, physics, and ecology, create a considerable opportunity for this rich information to be exploited by ML. Consequently, ML may contribute to understanding the past, present, and future traffic trends in the High North, from which derived knowledge may be used to develop proper emergency preparedness and SAR strategies.

1.2 Research objectives

The overall objective of this master's thesis is to investigate how to utilize ML and AIS data for identifying activity trends in the Arctic and how these trends, for different types of vessels, correlate with factors related to the Arctic environment, ecology, and the climate changes. This objective is further explained by the following research questions being addressed:

1. How to develop a supervised ML framework for predicting vessel presence and vessel density in the Arctic waters by time and space?
2. How do ecological and environmental factors in the Arctic affect activity presence and the level of vessel density?
3. Is supervised ML a reliable approach for activity predictions in the Arctic?

1.3 Scope and limitations

To achieve the objective of the thesis by available data and hardware resources, the definition of the Arctic region in this study is narrowed down to cover an area spanning from the northern coast of Norway to above the Svalbard island in the Barents sea. The relevant period is from January 2015 to May 2021, and the different maritime industries investigated are fishing, commercial shipping, and tourism.

The Barents Sea was selected according to its fast changes in climate and vessel traffic. Warm Atlantic water enters the Barents Sea from the southwest, and fresh, cold Arctic water

enters from the north, creating a separating feature known as the Barents Sea Polar Front. The polar front has substantial variations in temperatures and salinity, which may impact different types of maritime activities. Moreover, The Barents Sea is one of the world's most productive oceanic areas and has been important for the Norwegian fishing industry for several decades (Arctic Council 2021). Warmer seas will lead to additional species entering the region, increased fish stock sizes, and prolonged productive seasons. Hence, an increasing fishing trend is expected (Norwegian Polar Institute n.d.). In addition, Svalbard has long been a popular destination for tourism activities. Tourists are attracted to sea ice and the exotic wildlife existing further north, which has resulted in a significant increase in the number of visits in recent years (Svalbard Museum n.d.).

The work of this thesis relies on the quality and extent of data. As data quality has a major impact on an ML model's prediction performance, an essential part of the thesis is based on analyzing and understanding the quality and coverage of the different data sources applied. In addition, AIS data do not cover all vessel traffic, such as smaller vessels, leisure crafts, and government and military vessels. Hence, the experimental results underrepresent the actual vessels operating in the Arctic.

1.4 Contributions

The main contribution of this thesis is to provide an applicable proof of concept where ML is applied to Arctic environmental and physical features for activity prediction. As such, the thesis suggests an alternative approach to traditional statistical methods for identifying potential responses to the multiple changes affecting the region.

Projections of activity trends are useful for the global maritime industry and strategic planning by governments to understand and assess future spatial and temporal ranges of the Arctic operations. The proposed work aims to serve as a foundation for guidance related to construction, management, and maritime Arctic infrastructure in order to mitigate Arctic risks and improve emergency preparedness and SAR strategies.

1.5 Thesis outline

The remaining chapters of this thesis are structured as follows:

- Chapter 2 aims to provide the reader with insight into relevant theory, including background on the Arctic region and its associated risk, as well as technical and conceptional foundations on AIS and ML.

- Chapter 3 briefly presents the individual data sources used for allocating relevant data that represent the final predictors and prediction targets used in the proposed ML solution.
- Chapter 4 explains in detail the methodology applied in this project. This includes domain definitions, retrieval of relevant data, analysis of the data, and the ML model development process.
- Chapter 5 describes the experimental results from the proposed ML approach applied to two case studies with different activity viewpoints.
- Chapter 6 includes discussions related to the proposed solution, the validity of the given results, and its limitations.
- Chapter 7 concludes the work conducted and proposes further work.

Chapter 2

Theory Foundation

This chapter provides an overview of existing research from the literature combined with explanations of theoretical foundations in a broad context that serve as background for the work conducted in this thesis. Relevant papers and publications were retrieved directly from references and the search engines Engineering Village and Scopus. The keywords applied in this study were *risk, safety activity, mapping, prediction, Arctic, machine learning, maritime, marine* and *AIS*. Sections 2.1 through 2.4 cover information related to the Arctic region, including common sea routes, characteristic risk factors, established definitions and regulations, and an overview of the currently evolving activity trend. Sections 2.5 through 2.7 consider relevant ML definitions, taxonomies and approaches, and describe the processes from data gathering to ML model development. Finally, Section 2.7 presents previous studies on ML applications related to operational vessel activity.

2.1 Arctic sea routes

The original founders of the Arctic were indigenous peoples searching for supplies, food, and areas to settle. Western marine transport entered the region in the 1500s, initially motivated by finding alternative sea routes connecting Europe and Asia. Since then, several Arctic voyages have taken place, and marine shipping has advanced in vessel construction, infrastructure, governance, and improved crew training (Arctic Council 2009). As a result, there are three principal shipping routes in the Arctic that connect the Atlantic to the Pacific:

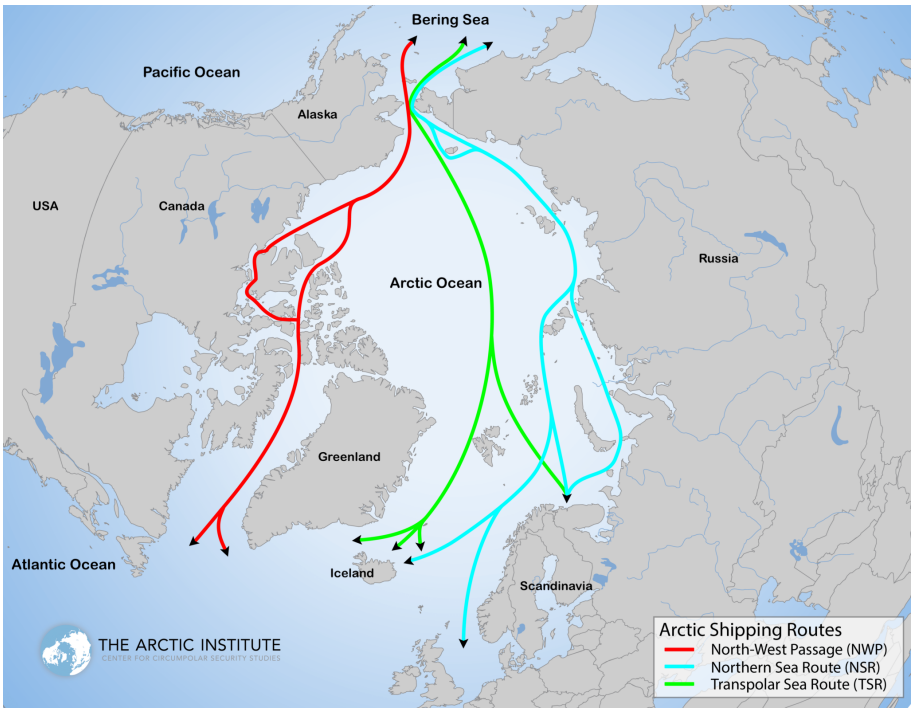


Figure 2.1: The three major Arctic shipping routes. Illustration adapted from The Arctic Institute (2015)

The Northwest Passage (NWP) traverses the Arctic Ocean, following the North American coast. The passage was completed for the first time by Roald Amundsen in 1906, and the first commercial transit of success occurred in 2013 by a bulk carrier. Some publications predict ice-free summers by 2050, hence 100% accessibility during July, which will augment the expansion of transits through the sea route (Marsh Risk Management Research 2014; Stephenson, Smith, and Brigham 2013; Whiteman et al. 2021). However, channels suitable for large vessels are expected to have challenging ice conditions for many years ahead (Peters et al. 2011).

The Northern Sea Route (NSR) connects the Bering Strait to the European waters along the Russian Arctic coastline, following the edge of the Norwegian Barents Sea. It reduces the distance between the Asian and European markets by up to 40% compared to the Suez Canal (M. Liu and Kronbak 2010; Stephenson, Smith, and Brigham 2013). Along with the melting sea ice, the route has become more attractive commercially and internationally. It is expected to be the preferred route among the three major Arctic routes toward future economic activity (Marsh Risk Management Research 2014).

The Transpolar Sea Route (TSR) is the most direct trans-Arctic shipping route, which connects the Atlantic and Pacific oceans by passing through the North Pole. Currently, this route is mainly covered with ice. However, with even warmer temperatures in the future, it is expected that the route will offer distance savings and become a popular alternative voyage (Ocean Conservancy 2017).

2.2 Arctic risk

The Arctic Circle spans around the globe at 66°N. Many scientists specify the Arctic as the area north of this circle where the sun does not rise or set at least once a year. Furthermore, organizations and institutions have other definitions according to vegetation, political considerations, or average summer temperature, such as below 10°C (NSIDC 2020). For example, the IMO identifies the Arctic based on their Polar guidelines, which spans around the globe at 60°N with deviations around the waters surrounding Iceland and the Norwegian mainland (IMO 2017). Although there are numerous definitions of the boundaries composing the Arctic area, many of them share that they are based on unique ecological, environmental, and physical characteristics, of which some are not to be found anywhere else in the world.

Moreover, the same characteristics make up a complex risk picture and are often referred to as the Arctic risk influencing factors (RIFs). A RIF is defined as "an aspect, such as an event or condition of a system or an activity, that affects the risk level of this system or activity" (Øien 2000). On the other hand, *risk* is an expression of the expected amount of harm resulting from an event occurrence and is traditionally calculated as the product of event occurrence multiplied by a measure of the event's corresponding consequences on people and the environment (Marchenko et al. 2018). In 2017, the Norwegian Coastal Administration (NCA) initiated the development of a guideline for marine risk assessment in the Arctic (EPPR 2018). The project involved identifying the major Arctic RIFs. Table 2.1 summarizes these RIFs and how they individually are sources of increased operational event occurrence and further contribute to negative impacts from the event's effects.

Arctic RIF	**Incident impact (causes)**	**Consequence impact (effects)**
Sea ice	• Collisions with ice • Vessels being stuck in ice • Changes in vessel stability • Deviations from planned route	• Hindrance of evacuation • Structural damage causing pollution
Poor visibility	• Hindrance of visual identification of other objects at sea • Interruptions of human performance • Deviations from planned route	• Challenges in SAR operations • Challenges in cleanup of oil spills
Low temperature	• Icing of equipment • Freezing of fluid • Human discomfort	• Poor survivability and emergency performance in cold waters
Remoteness	• Poor contact with shore • Insufficient information of surroundings	• Hindrance of evacuation • Nearby assistance less available • Delays in SAR operations
Human experience	• Poor Arctic knowledge and confidence	• Lack of emergency equipment • Lack of mitigation measures
Violent weather conditions	• Uncertain weather forecasts • Maneuver challenges	• Escalation in weather conditions • Challenges in SAR operations
Emissions spills and ballast water pollution		• Mortality to marine ecosystems • Invasive species • Global warming
Environmental and ecological vulnerability		• Long-term restoration • Species of extinction

Table 2.1: Arctic RIFs. Information based on EPPR (2018)

The RIFs associated with temperature conditions and sea ice have caused the Arctic to be inaccessible for centuries. Consequently, the probability of accidents in the region has been considered low. On the other hand, the consequence factor in the risk expression has always contributed to a high-risk product. This is related to the RIFs presented below the double line in Table 2.1, which concern the fragile Arctic environment and the polar ecosystem's poor resilience to human disturbances (Ocean Conservancy 2017; Arctic Portal n.d.). If an incident arises, emergency performance and cleanup operations from emission spills will be challenging due to the harsh Arctic conditions and the lack of response capabilities. For example, oil spills in the Arctic Oceans might remain in the ice for several decades due to the near-zero temperatures (Lahn and Emmerson 2012). As such, incident scenarios can have catastrophic and long-lasting impacts on the Arctic compared to southern climates.

In recent times, however, global warming has affected the RIFs that previously prevented maritime traffic from entering the Arctic. According to new transportation routes and discoveries of resources, there is a growing interest in entering the region within several industries (PAME 2021a; PAME 2021c). Such a rapid increase in the number of vessels leads to similar growth in incidents. Consequently, higher vessel densities make the final risk

measure more dominated in terms of event probability and not just the outcome of the event. This emphasizes the strong connection between activity presence and operational risk (Arctic Council 2009).

2.3 Rules and regulations

Traditionally, the Arctic oceans have not been under any specific authority. However, since the 1700s, countries have tried to claim parts of the oceans that border their coasts. As a result, the United Nations Convention on the Law of the Sea (UNCLOS) established various maritime boundaries of the Earth's water which encompass maritime limits and zones of rights over marine and biological resources (Sea Around Us 2015). Still, the activities taking place in the Arctic are not governed by any single legal appliance, body, or regime. Instead, the governance is conducted by a composition of international regulations and cooperation among the eight Arctic states: Canada, Denmark, Finland, Iceland, Norway, Russia, Sweden, and The United States (AWI 2020).

2.3.1 International Maritime Organization

The International Maritime Organization (IMO) is an agency under the United Nations responsible for shipping safety and the prevention of marine pollution. The IMO is the source of global mandatory and voluntary standards and regulations at sea. Their work has developed technically and legally according to increased activity during the last five decades (FN-sambandet n.d.).

The International Convention for the Prevention of Pollution from Ships (MARPOL) and the International Convention for the Safety of Life At Sea (SOLAS) are two fundamental IMO conventions established to respond to the world becoming more aware of environmental harm caused by the growing skipping industry. While MARPOL addresses marine and atmospheric pollution caused by ship operations and accidents, SOLAS works toward safety standards regarding constructions and procedures of merchant ships (United Nations 2017). According to the recent growth in vessel activities around the Poles, the IMO adopted the International Code for Ships Operating in Polar Waters (Polar Code), which is mandatory under both MARPOL and SOLAS. The code entered into force in 2017 and includes ship-specific requirements and safety and environmental regulations for vessels with intended operations within the Polar regions (IMO 2017).

2.3.2 Arctic Council

The Arctic Council is an intergovernmental forum that promotes and works for sustainable development and environmental protection of the Arctic region. The forum includes the members of the eight Arctic States and Indigenous Permanent Participant organizations (EPPR 2018; Arctic Portal n.d.). Their activities are conducted through six working groups, each working toward a particular objective. The most relevant working group for this thesis is the Protection of the Arctic Marine Environment (PAME) group. Their activities are carried out through a bi-annual work plan and include regional and circumpolar guidelines and sea and land-based activities aiming to protect the Arctic marine environment. In 2019, PAME introduced the Arctic Ship Traffic Data (ASTD) system, which is a repository of reliable, accurate, and up-to-date information on vessel activities specifically focused within the Arctic region (*PAME - Arctic Ship Traffic Data* 2022). The system aims to support analysis within the Arctic Council's working groups for assessing changes, monitoring trends, and reducing knowledge gaps related to Arctic traffic in light of the changing environmental conditions (ASTD PAME 2020).

2.3.3 Automatic Identification System

AIS is a collaborative system that enables a vessel to automatically transmit and communicate information to shore stations, satellites, and other ships in its neighborhood. The frequency of transmitted signals varies from two seconds to three minutes, depending on the ship's speed and position, where higher speeds correspond to higher frequencies (ASTD PAME 2021). The system was developed to facilitate collision avoidance and is widely used for maritime situation awareness, surveillance, and pollution monitoring. It was initially standardized by the International Telecommunication Union (ITU) before the IMO later adapted it to include additional ship-specific and safety-related information such as position, current timestamp, speed, course, and static vessel characteristics. Each vessel is distinguished by its Maritime Mobile Service Identity (MMSI) number (Zhong, Song, and Yang 2019).

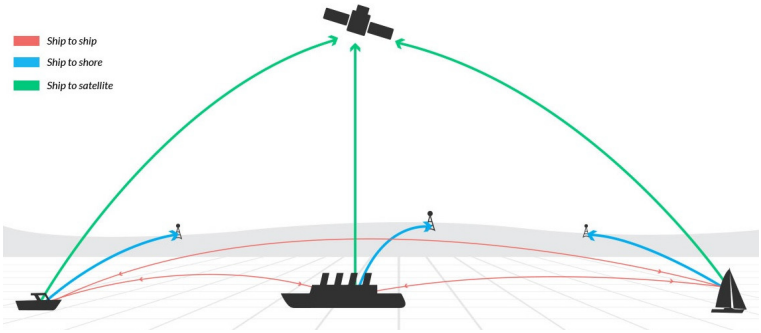


Figure 2.2: Automatic transmissions of AIS signals from ship to ship, to shore and to satellite. Illustration retrieved from NATO Shipping Centre (2021)

While the global fleets have carried AIS since the 1990s, the system has received more attention during the last decade according to new regulations and technological development. The AIS standards have evolved since SOLAS adopted the first requirement for AIS carriage in 2004. All vessels of 300 gross tonnages and upwards engaged on international voyages, all cargo ships of 500 gross tonnages and upwards, as well as passenger ships irrespective of size, are required to carry an AIS transceiver (IMO n.d.). In 2014, AIS became compulsory for all fishing vessels above 15 meters in length within the EU fleet (Global Fishing Watch n.d.).

The increased and global use of AIS systems facilitates digitization in the maritime field, as the high volumes of real-time information constitute a rich data source. However, AIS data do not cover all ship traffic, as leisure crafts, smaller vessels, and government and military vessels are not required to carry the transmitters. Additionally, several factors may affect the signals, resulting in gaps in the data. These include technical failure, manipulations of signals, data noise, satellite coverage, and incorrect installation (ASTD PAME 2021).

2.4 Related work on Arctic activity trends

New technology trends combined with vanishing sea ice facilitate industrial activities in the High North, including oil and gas extraction, commercial fisheries, and tourism. It is believed that the Arctic areas contain spots of undiscovered oil and large quantities of minerals, potentially leading to more offshore energy development and mining operations, resulting in increased trans-Arctic shipments (Peters et al. 2011). The Arctic's oceans are also known for their productive fishing grounds, of which the Barents sea supports the world's largest cod stock (Loe et al. 2014). Moreover, the exotic wildlife, culture, and pristine Arctic landscape make the region a popular tourist destination. Consequently, the individual industrial sectors work as catalysts toward increased vessel traffic, which several studies have lightened.

After the ASTD was launched in 2019, the PAME working group performed several statistical analyses on vessel traffic in the Arctic. PAME (2021a) analyzed AIS data from 2013 to 2019 with a specific focus on passenger vessels and cruise ships within the Polar Code definition of the Arctic. Their statistics showed that the overall unique ship count had increased by a factor of 22, but the relative increase in the tourism segment was small. Cargo ships, on the other side, had a much higher increase according to PAME (2021b) which revealed a 160% relative increase by unique ship count of bulk carriers. PAME (2021c) investigated shipping trends in the NWP by different measurements of activity volume, including unique ship count and distance sailed. Both measures revealed an increase in activity and confirmed that bulk carriers had the largest traffic increase. Similarly, M. Liu and Kronbak (2010) and Stephenson, Smith, and Brigham (2013) designated the NSR as the future transit between Asia and Europe. However, due to the lack of economic centers along the route, risks associated with the Arctic conditions, and the trade-off between time savings and potential costs from ice breakers, there is still doubt as to whether the passage will substitute the Suez Canal significantly (Jensen and Paglia 2021; M. Liu and Kronbak 2010; Peters et al. 2011).

Although multiple analyses of the Arctic traffic volume have proven a significant traffic growth during the last decade, especially within the cargo shipping industry, the future projections still vary due to uncertainties related to the navigation season, development of regulations, expansion of resources, and variable sea ice. While future changes in regulatory frameworks or shifts in political spheres are challenging to foresee, various studies have applied climate models to project future maritime access to the Arctic. Such models are programmed to forecast or recreate past climate conditions, including atmosphere, ocean, sea ice, and land surface models. Hence, they may facilitate understanding future effects and impacts of global warming (NCAS n.d.). One example is the Arctic Transport Accessibility framework developed by Stephenson, Smith, and Agnew (2011), which integrates climate modeling projections for quantifying the navigation access to the Arctic oceans, including air temperature, sea-ice scenarios, topography, hydrotherapy, infrastructure, and human settlements. Their model revealed that the NSR and the TSR will become 100% accessible for vessels with limited icebreaking by mid-century.

Other studies have applied the output from climate models to investigate the subsequent response in vessel traffic. Peters et al. (2011) developed emission estimations of Arctic shipping related to petroleum activities in 2030 and 2050. Climate models estimated future ice coverage, and estimations of transit shipping were based on a cost-benefit analysis comparing alternative routes through the Suez Canal. Lastly, a global energy market model was used to predict

future petroleum activities. Their results showed that a decrease in sea ice coverage led to rapid emissions growth from trans-shipments. Fauchald et al. (2021) performed an estimate of how fishing activities respond to climate model projections of warmer Arctic climate and diminishing sea ice. They developed a statistical model for predicting the presence of trawling vessels using sea surface temperature, sea ice concentration, and bathymetry data as explanatory variables. Their results proved that trawling activities expanded rapidly by inter-annual sea ice loss.

Ultimately, it is evident from the investigated literature that activity, either measured by vessel density, fuel consumption, or emissions, is expected to increase with the changing climate variables in the Arctic, of which sea ice is considered the most significant obstruction to vessel navigation.

2.5 Supervised machine learning

ML is a subfield within Artificial Intelligence (AI) that describes computer algorithms that automatically learn patterns from past experiences provided through vast amounts of training data. The overall goal of ML is to develop computer systems that are able to use such patterns and solve complex problems accordingly. Numerous different ML algorithms can be applied to different types of problems, of which the target of prediction can either be numerical or categorical.

Supervised learning is a subcategory of ML which covers algorithms that are provided with a correctly labeled training data set of input-output pairs and learns a function that maps the input to the output. The input corresponds to the data features, or *predictors*, and the output corresponds to the *target of prediction*. The function to be generated is usually denoted as the *hypothesis* (h) which is the ML model's approximation of the true unknown function (f) that generates the output (y_i) from the input (x_i) (Russel and Norvig 2020b). As such, the supervised learning task can be formulated mathematically as follows:

Given $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$,
where $y_i = f(x_i)$, $N \in \mathbb{R}$, $i \in N$,
find $h \approx f$

The performance of the supervised model is measured by testing the hypothesis on new observations. This is done by applying it to a test set that is distinguished from the original training set before its exposure to the model learning process.

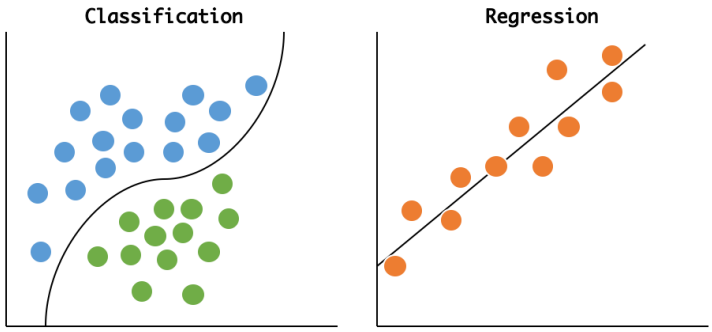


Figure 2.3: Illustration of 1) a binary supervised classification problem and 2) a linear regression problem

Furthermore, supervised ML can be categorized as a classification or a regression problem. The former uses a model function that separates the data into multiple classes and predicts a categorical label within this finite set of classes. The latter uses a model function that finds the correlation between numerical input values and quantifies a number from the continuous quantity (Goodfellow, Bengio, and Courville 2016). This difference is illustrated in Figure 2.3, which shows a classification task by two classes and a regression problem of linear correlation. The classification task is considered as *binary*, as it only has two possible output values. However, a classification task can also involve three or more potential outcomes, commonly known as a *multiclass* problem.

Dependent on the target definition and the desired model performance and execution speed, several supervised learning approaches may be suitable candidates (Caruana and Niculescu-Mizil 2006). The algorithms are distinguished by their underlying structures of organizing the input data and searching through the space of hypothesis (Mitchell 1997). Some of the most commonly used methods found in the literature search related to this work will be described in the following.

Decision Tree learning is a practical and widely used method for inductive learning. Decision Trees map all possible solutions from a preliminary question asked. They are structured graphs, creating the supervised patterns by a flowchart structure through branches and nodes. The nodes represent the data features, and the connecting branches correspond to the features' values. At each hierarchical level of the tree, a condition on a specific feature is asked, of which result follows the value through the corresponding branch, moving to the next

level, i.e., feature. The final level constitutes the leaf nodes which specify the final prediction (Mitchell 1997).

Random Forest (RF) was first introduced by Breiman (2001), and is an efficient bagging-based ensemble learning method that combines multiple decision trees, i.e., a forest, to improve performance. Instead of relying on the output from one single decision tree, many such trees are built in parallel from randomly selected subsets of the original data set. This increases the diversity of the model. The aggregated result from the decision trees yields the final result as seen in Figure 2.4.

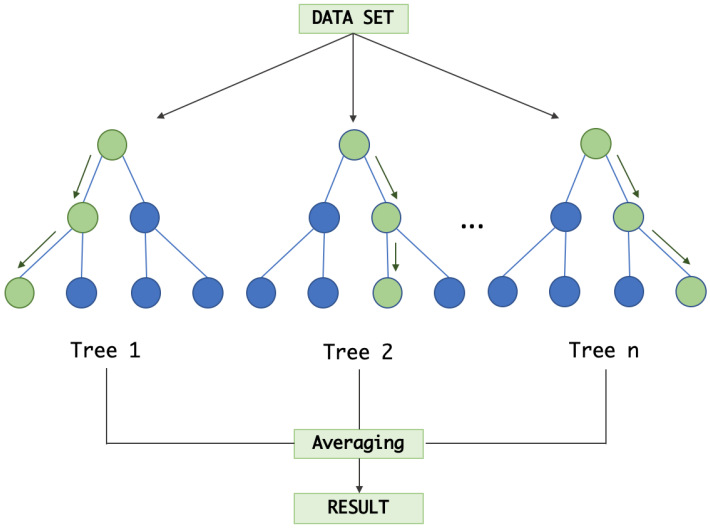


Figure 2.4: The RF learning method. The final result is aggregated from independently built tree predictors

Gradient Boosted Tree is another ensemble method based upon a set of different decision trees. However, it differs from RFs by how the model is trained and the final aggregation of the output. As illustrated in Figure 2.5, the trees are built gradually and sequentially, where the residual between one model and the previous, i.e., the delta between the correct value and the prediction, is used as further input to enhance performance. As such, this boosting mechanism aims to minimize the predecessor's error. Compared to RF, Gradient Boosted Trees aggregate the final result during the learning process and not after. It is one of the most commonly used tree models by researchers due to its simplicity and proven high performance (Chen and Guestrin 2016).

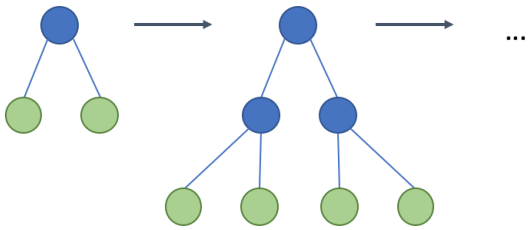


Figure 2.5: The Gradient Boosted method with level-wise tree growth

Support Vector Machine (SVM) represents the training data as spatial points and builds a model of N-dimensional boundaries where N i is the number of features in the data. The boundaries are used for classifying new input labels based on their measured distances in space (Russel and Norvig 2020a). As illustrated in the binary problem in Figure 2.6, the objective is to find the hyperplane having the maximum distance between points of different target values such that future predictions may be obtained with more confidence (Chatzikokolakis et al. 2019).

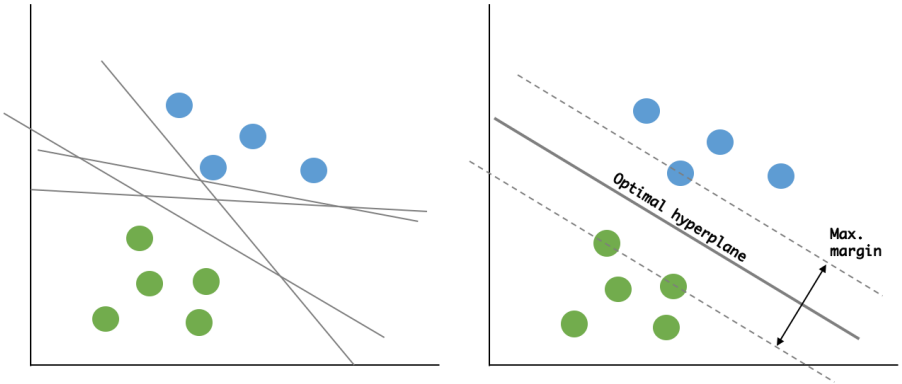


Figure 2.6: SVM of two classes

Artificial Neural Nets (ANNs) are inspired by the biological neurons of the human brain and aim to approximate functions in complex data. As seen in 2.7, they are structured by several interconnected units represented in layers that sequentially transfer data from the input layer to the final prediction in the output layer. The edges that connect the layers are

associated with weights that are adjusted during the learning process. These weights define the strength of data signal from one layer unit to another (Mitchell 1997).

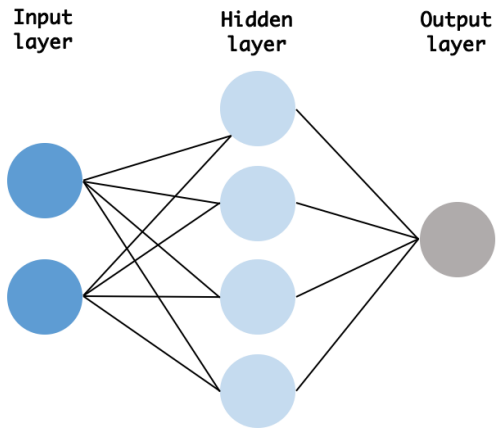


Figure 2.7: ANN with one hidden layer

2.6 Machine learning workflow

The ML workflow serves as the basis for the methodology of this project. It describes the iterative process of developing the ML model and consists of an end-to-end set of standardized steps. Firstly, the project domain and the objectives that the model tends to achieve must be defined. This includes defining the specific target of prediction and identifying desired predictors from potential data sources. From there, the workflow is divided into 1) pre-processing the data, which includes data gathering, data cleaning, exploratory data analysis (EDA), and feature engineering, and 2) model development, which includes training, testing, and model evaluation (Kuhn and Johnson 2019). Each step should be completed sequentially in a loop, as given in Figure 2.8.

The aim of the workflow is to understand and prepare the raw data for the intended application in order to obtain the most optimal prediction performance on new data. In fact, the modeling technique itself is typically a small part of the model development process. A developer rule is that only 20% of the time should be spent on model training and predictions, and the remaining 80% should be spent on data pre-processing (Khalitov 2021a).

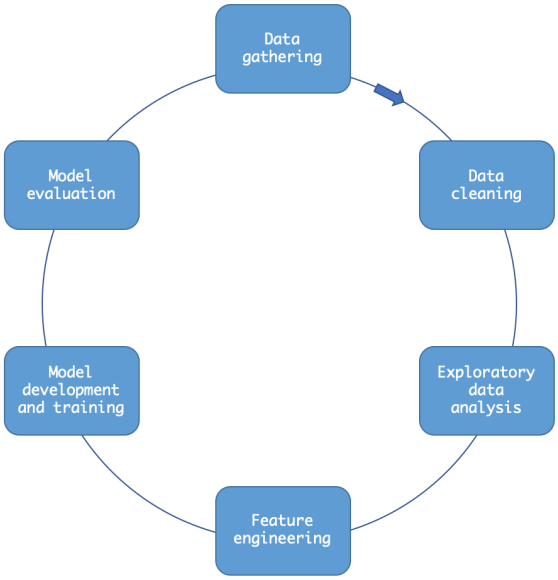


Figure 2.8: The ML workflow

2.6.1 Data gathering and data cleaning

Data collection is commonly defined as one of the major challenges in ML. The data acquired form the basis of the ML workflow and the model's final usefulness and plays a vital role in any ML use case. Therefore, it is essential to search thoroughly through data providers and databases to allocate the predictors that best suit the purpose of the model (Yellenki 2020). Although some data sets are open source, most providers require account registration or special permissions due to data confidentiality concerns. Consequently, much of the work lies in this step.

The features of the raw data collected are typically inconsistent and erroneous and must be prepared before being exposed to further transformations through feature engineering. The cleaning process includes joining data from multiple sources, extracting the data by desired temporal and graphical boundaries, identifying and treating missing values, nonsensical values, duplicates and improperly formatted data, and so on. The aim is to maximize the data quality without necessarily losing information (Khalitov 2021b). The cleaning techniques found relevant in this project will be described in the following.

Data redundancy and irrelevance refer to unnecessary information for enhanced prediction. Sensor data and data collected from multiple different sources may contain duplicate observations. Duplicates are data records representing the same entity, such as

similar positional signals from AIS being captured twice. On the other hand, irrelevant data are data that do not fit into the purpose of the model. Both duplicates and irrelevant observations have in common that they increase the ML model's computational cost without providing additional valuable information (Kuhn and Johnson 2019). Therefore, it is beneficial to restrict the data to relevant observations and remove duplicates prior to other modifications.

Noise and Outliers are ML concepts describing data discrepancies. Real-world data most certainly contain errors or mislabeled samples, for example, caused by unreliable collection instruments. In ML, these errors are called noise and are either errors in the predictors' values or incorrect target values. On the other hand, outliers may contain erroneous data, but not necessarily. These refer to unrepresentative data and are abnormal labels that differ from the remaining data. Hence, there is no direct link between outliers and noise as outliers can contain valuable information about the underlying system (MIT 2021). This has, for example, been exploited by several studies in anomaly detection in AIS tracks, potentially indicating criminal activities or accidents (Wolsing et al. 2022; Li et al. 2017; Goerlandt et al. 2017). However, outliers in environmental data evolving in time and space are typically errors due to inconsistency and should be either smoothed or removed.

Missing values refer to lacking information in the data, and are typically encoded as *NaN* (Not a Number) values. Such values may lead to a biased ML model, and many algorithms even require complete predictors (Kuhn and Johnson 2019). Bias occurs when there are systematic errors in the data which affect the learning process' assumptions and learning patterns, and skew the predictions towards a wrong idea (Giffen, Herhausen, and Fahse 2022). In order to deal with the missing values, it is important to understand why the data are missing. Some typical reasons could be improper maintenance of past data, information not provided intentionally, or failure in recording observations due to technical or human error. Hence, missing values are formally divided into three categories:

- Missing completely at random (MCAR): the missing values are neither related to the other missing values nor the existing values.
- Missing at random (MAR): the missing values are related to the given observations but are unrelated to the remaining missing data.
- Not missing at random (MNAR): the missing values are related to other missing data.

Depending on the number of values missing and whether the missing information is MCAR, MAR, or MNAR, there exist various alternatives of manipulating strategies. As the missingness

within MNAR values is connected, an option is to encode these to new values. The easiest solution for cases concerning MCAR and MAR is to exclude the NaN values completely, either by deleting the affected records or removing a feature of many missing values completely. However, this can lead to a significant loss of valuable information. Hence, imputation, or filling-in, is usually desirable when there are many records of missing values for a relatively small number of features. Missing values can effectively be imputed by a specific metric, such as the most frequent value, the mean, or the median of the already represented data. In recent years, however, more advanced imputation methods have been introduced, which involve applying supervised models on the represented data using the missing values as prediction targets (X. Liu 2016).

2.6.2 Feature engineering and exploratory data analysis

The predictors, or features, in the gathered data may be represented in a way that makes it difficult for the ML models to achieve good performance. Feature engineering is about reworking and adjusting the predictors to strengthen the relationships between the predictors and the target values. For example, this includes feature transformations, interacting multiple features together, or constructing re-representations of features (Kuhn and Johnson 2019).

Determining and selecting meaningful predictors is an additional important part of feature engineering, as features of non-informative values decrease the model's interpretability, training speed, and performance. An individual feature's usefulness for target prediction is evaluated by its degree of information gain. Higher information gain improves the knowledge of the model and reduces the need of information for further training on the remaining data. In decision trees, the feature importance is calculated by the Gini impurity metric, which measures the quality of a split by the likelihood of new observations being misclassified after the split. This likelihood is calculated by the number of times a randomly chosen sample from the split-subset would be incorrectly classified. Hence, the lower the impurity, the higher the information gain (Mitchell 1997).

The best reworking approaches for a given ML problem are typically unknown, and many alternatives could be searched for in order to find the most appropriate ones (Kuhn and Johnson 2019). Therefore, EDA is necessary to gain insight into the data. EDA refers to investigating the data using visual techniques and statistical graphics. The aim is to understand the underlying feature distributions and their interplay, test assumptions, and discover data anomalies. The output from the EDA is further used as a basis for any potential

re-representations and the final selection of meaningful predictors to use as input prior to modeling (Tukey 1977). As such, feature engineering and EDA go hand in hand.

2.6.3 Model development

Once the data are processed, the data set must be split into a training set and a test set. The model uses the training set for processing information and mapping predictor values to their target values. The test set is kept unavailable for the model during model training and is used to evaluate the model performance after training by assessing the accuracy of the predictions compared to unseen truth values.

The model can be refined during model training until an acceptable level of accuracy is achieved. Many ML models are provided with a set of predefined settings, given as *hyperparameters*, which can be configured explicitly according to the training data and the ML problem's objective. Hyperparameters control the model's trade-offs and decisions during the learning process, influencing the model's time complexity and quality. Tree depth, number of trees, and the fraction of features used to build a tree are typical examples of hyperparameters for tree-based models (Kuhn and Johnson 2019; Banarjee 2020). Developing the ML model involves finding optimal hyperparameters that leverage the model's maximum power. k-fold cross-validation is a technique that can be used to assess model performance before exposure to new observations and choose optimal hyperparameters from predefined sets of configurations (Russel and Norvig 2020a). In k-fold cross-validation, the training data is split into k number of subsets (folds), and the model is trained and evaluated in an iterative process. The process goes on for k times, each time training the model on $k-1$ folds and evaluating it with the last kth fold. An aggregation of the individual evaluations determines the final model performance (Kuhn and Johnson 2019). The process is summarized in Figure 2.9, which shows cross-validation by five folds.

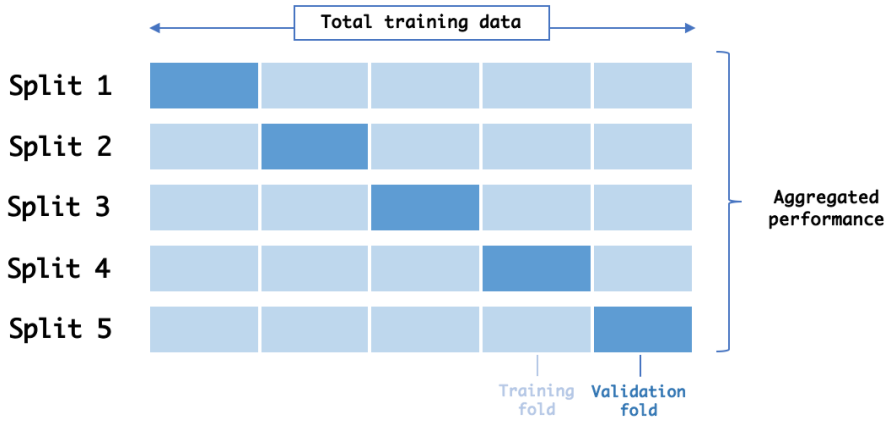


Figure 2.9: 5-fold Cross Validation

The choice of k depends on the ML problem and the training data size, of which common values are 3, 5, and 10. A higher number of folds means a larger training set relative to the test set, resulting in a lower prediction error as the model is exposed to more input data. However, choosing a smaller fold value is often beneficial as it requires less execution time.

2.6.4 Model evaluation

An unbiased evaluation of an ML model's learning performance is estimated by applying the model to new observations, either provided by validation sets *during* model training or a hold-out test set for a final evaluation *after* training. The model can be assessed by various performance metrics, some of which are more appropriate than others depending on the model domain, the algorithm used, and the prediction target. Furthermore, the metrics measure model trade-offs differently so that they may perform more optimally by one metric than another. Therefore, it is recommended to use several metrics to evaluate model (Caruana and Niculescu-Mizil 2006). Some of the most commonly used metrics for classification tasks are described in the following.

Logarithmic loss is a measure based on the probability of a sample being classified under a certain class. It is typically used by gradient boosting methods and ANNs, whose objectives are to minimize an error function. It describes how close the probability is to the actual value and, implicitly, how confidently the model predicts its output. Hence, the smaller the logarithmic loss, the higher the model accuracy (Pedregosa et al. 2011).

Accuracy is the simplest metric and yields the ratio of correctly predicted records to the total number of predictions made, as expressed in 2.1. It efficiently estimates the overall model performance and works well on balanced data. However, as further elaborated in Section 2.8.2, most real-life ML cases concern imbalanced data classes that preferably should be subject to individual evaluation.

$$Accuracy = \frac{\#correct\ predictions}{\#predictions} \tag{2.1}$$

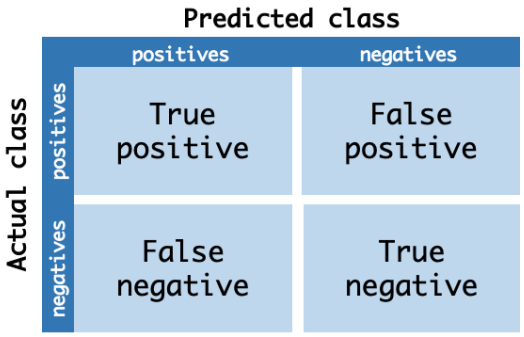


Figure 2.10: Confusion matrix of a binary classification problem

Confusion matrix is a tabular representation of the model performance within each class of a classification problem. Figure 2.10 presents such a matrix for a binary problem, where each column corresponds to a predicted class, and the rows correspond to the true classes. The matrix illustrates four possible outcomes of class predictions. In a binary task, in which the target value is either *yes* or *no*, the aim is to predict yes when the true target class is yes and, similarly, no when the target is no. Hence, the model obtains correct predictions by true positives (TP) and true negatives (TN), respectively. The opposite outcomes occur when the model predicts a class value that contradicts the true class value. In this case, the predicted class is either no while the actual class is yes or the predicted class is yes while the actual class is no, which yields false positives (FP) and false negatives (FN), respectively. An optimal learning algorithm would result in a confusion matrix of entries only along the main diagonal, meaning that all predictions are TP and TN, that is, correctly estimated (Kotu and Deshpande 2015). Together with the confusion matrix come three additional metrics which enable investigating performance by each prediction class individually:

- **Precision** is the ratio of true positives, or relevant cases, to all positive predictions made, as given in 2.2. It yields how many true values the model correctly predicted out of all values the model predicted as true.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

- **Recall** is the ratio of true positives, or relevant cases, to the actual true classes, as given in 2.3. It is a measure of sensitivity by the model's ability to capture all cases that actually are relevant.

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

- **F1-score** is a weighted combination of both precision and recall. It is less interpretable than accuracy, however considers both false negatives and false positives which makes it more suitable for evaluating model performance from uneven data distributions.

$$F1 - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{2.4}$$

2.7 Previous activity estimates by machine learning

The increased use of AIS transceivers has led to more accurate and easily accessible vessel traffic information, facilitating knowledge extraction on activity trends using data-driven methods. This is exemplified by several studies from the literature that combine ML with historical AIS from around the world. The existing AIS-based ML applications include vessel type classification, anomaly detection, and prediction of ship characteristics and trajectories (Wolsing et al. 2022; Li et al. 2017; Mazzarella et al. 2014). Although there is a lack of such studies with an Arctic-specific focus, the existing research may provide good indicators on which approaches are suitable for these vast amounts of sensor-based data.

Chatzikokolakis et al. (2019) applied Decision Trees, RFs and Gradient Boosted Trees with a large volume of AIS data to detect SAR activities. The study aimed to evaluate the learning approaches' ability to identify trajectories of SAR vessels. The data were localized from diverse areas in need of such activity due to migrants reaching for Europe, including the Central, Eastern, and Mediterranean routes. 5-fold cross-validation was used to determine the number of trees and their corresponding depth. Their proposed approach proved to be generic in different areas of interest. Gradient Boosted Trees performed with the best accuracy by all evaluation metrics but had higher computational demands. Hence, the authors concluded with RF as the outperforming model according to efficiency and accuracy. Nguyen et al. (2018) jointly addressed anomaly detection, vessel type identification, and trajectory

reconstruction by a multitask ANN approach. The key component of their framework was an embedded block that converted input streams of AIS data to regularly sampled data. Blocks of higher levels were task-specific submodels for either anomaly detection, vessel classification or trajectory reconstruction. The submodels were tested using AIS data from the Gulf of Mexico from January to March 2014. To detect anomalies, the authors constructed track divergences and circle-shaped patterns in places where these are not expected. Compared to state-of-the-art models for similar problems, the ANN approach detected similar abnormal patterns and slightly improved classification performance by an F1-score of 87.72%.

Multiple studies have applied ML with AIS data from fishing vessels in order to assist the management of fisheries concerning overfishing and illegal fishing behavior. One example is the work of Jiang et al. (2016), which used ANNs, SVMs, and RFs to detect fishing activity from AIS data. Their target was defined by a binary classification problem for the prediction of fishing vessel presence or absence. Due to imbalanced data in favor of the majority class, the training data were re-sampled. The results were evaluated by various metrics, which revealed improved prediction by all three ML approaches after re-sampling. Additionally, ANNs performed at least as well as SVMs and RFs.

Other studies have addressed inference of vessel types in general, aiming to improve Maritime Situational Awareness at sea. Kraus, Mohrdieck, and Schwenker (2018) collected AIS data covering vessel positions from the German Bight for distinguishing between fishing, passenger, cargo, and tanker vessels. The AIS data were used to extract other features, including vessel trajectories, the ratio of trajectory per vessel type, and distance to shore, which resulted in a training data set of both behavioral and spatial properties. RF was chosen for classification due to its trade-off between predictive capabilities and efficiency. Their results proved a satisfying performance with an overall accuracy of 97.51%. However, several tankers were misclassified as cargo vessels due to similarities in behavioral patterns. Similar results were obtained by Zhong, Song, and Yang (2019), which used global, real-world AIS data published by the National Defense University for vessel type classification between cargo vessels, tankers, and fishing vessels. The fishing vessels yielded the highest classification precision of 94.5%, but the model had difficulty distinguishing between cargo vessels and tankers.

As indicated above, several studies of ML predictions using AIS-based training data report promising results from tree-based learning models. This justifies the applicability of similar approaches with Arctic AIS data. However, most previous studies target vessel activity by vessel type or by behavior in terms of speed or trajectory patterns. These targets can be predicted from vessel-specific characteristics directly available from the AIS streams,

including size, gear, and positions. However, they are less dependent on factors describing the surroundings. For example, attributes representing vessel size and gear would be more explanatory compared to surrounding properties for answering whether a vessel is of type fishing or cargo. As such, the studies are limited to relying on features extracted from the AIS data only, without including other data sources as part of their training data. On the other hand, a quantitative measure of activity is an *aggregation* of the vessels transmitting the AIS records and is more dependent on external factors for prediction. Such factors include context-related information about the environment, ecology, and regulations affecting the activity quantities. For example, identifying the number of vessels traversing from one place to another would depend on the surroundings rather than the vessel's speed or gear. Although activity in this matter has been addressed by statistical methods, as presented in Section 2.4, there is a lack of attempts where ML approaches are applied in such cases.

2.8 Application challenges

Although supervised learning applications have emerged significantly and proven promising results in recent years, there are major challenges that may be faced when developing such applications. Some of the common challenges were experienced throughout the development of this thesis' proposed solution and will be described in the following.

2.8.1 Overfitting and underfitting

Overfitting occurs when the ML model obtains a small error on the training set but a large error on the test set. As a result, the model knows the training data well but cannot be applied to solve previously unseen problems. Such a situation usually happens when the model relies heavily on detailed patterns in the given observations that otherwise do not occur, Hence, the algorithm fails to *generalize* the learned behavior in future situations, i.e., adapt the learned patterns appropriately to unseen data. On the other hand, underfitting is when the model neither manages to fit well to the training data nor is able to generalize to new data. This happens when the model is not exposed to enough data to capture the underlying patterns (Goodfellow, Bengio, and Courville 2016).

While underfitting is easily detectable with appropriate performance metrics, overfitting is more challenging to handle as it may occur from multiple sources. Some well-known key takeaways for preventing overfitting involve reducing data complexity by properly pre-processing noise and redundancies and tuning hyperparameters to add randomness to make the training process more robust. In principle, the more computation allowed for training

increases the chance of a resulting complex model. Therefore, it is good practice to include early stopping in the ML model where the training process is stopped once the validation accuracy does not improve (Mitchell 1997).

2.8.2 Data imbalance

Data imbalance occurs when the target values are not equally represented. In regression problems, the target distribution is highly skewed towards a specific number, while classification imbalance is when most of the instances belong to the same class. As a result, the machine learning algorithm is exposed to more records of a particular target compared to others and becomes biased towards predicting the most represented target without actually performing any pattern analysis of the data (Kuhn and Johnson 2019). Consequently, the prediction output would mostly correspond to the true values of the target, yielding an apparent high accuracy. Evaluation by the accuracy metric on imbalanced data is therefore misleading as it does not consider the predictions of minority classes, which, in the case of imbalance, actually remark the predictive performance (Kotu and Deshpande 2015).

Imbalance occurrence is common in real-world domains where the aim is to detect rare but important cases, such as anomaly detection and text classification (Kotsiantis, Kanellopoulos, and Pintelas 2005). Although some ML models provide specific weights to even out the distributions during training, several solutions are proposed for manipulating the imbalances at data level. This includes various re-sampling approaches, which rely on either undersampling or oversampling. The former solution removes samples, i.e., records, from the most represented classes, while the latter adds more samples to the less represented classes. The different under- and oversampling approaches are distinguished by their choice of which samples to remove or how to generate new samples, respectively. For undersampling, the simplest approach is a random elimination of the majority classes until the ratio of classes is equalized. Correspondingly, random oversampling aims to balance the class distribution by randomly replicating samples from the minority classes (Lemaître, Nogueira, and Aridas 2017). Although these re-sampling methods solve the imbalance problem, they may act as sources of other pitfalls. For example, the removal of existing records of the majority classes causes a loss of potentially useful information for pattern recognition, while the creation of duplicates from the minority classes increases the chance of overfitting as the model is exposed to more similar information. Additionally, oversampling can result in an extensive computational task when working with large data sets (Kotsiantis, Kanellopoulos, and Pintelas 2005).

In order to decrease the likelihood of overfitting and prevent significant loss of information, there have been proposed alternative solutions to the random approaches and solutions which combine both oversampling and undersampling. One commonly used approach is the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002), which generates synthetic samples by interpolating from existing and closely related samples from the minority class. Consequently, duplicates are avoided, and the decision boundary is spread toward the space of the majority class. However, a high degree of imbalances may lead to many synthetic samples. Several researchers have therefore attempted to combine SMOTE with undersampling methods to even out the proportion of data removal against the generation of synthetic data (Sasada et al. 2020; Batista, Prati, and Monard 2004). One example is the SMOTETomek approach which combines oversamples by SMOTE with data removal of samples forming Tomek links. The concept of Tomek links was first introduced by Tomek (1976), and describes the linkage between two samples of opposite classes that are each other's nearest neighbors. These are used to exclude samples from the majority class close to minority class samples; hence, they create a more explicit decision boundary. As a result, the target classes become balanced by SMOTE generation followed by a clean-up based on Tomek Links, which has proven promising accuracy results (Z. Wang et al. 2019; Goel et al. 2013).

Chapter 3

Data and Technologies

This chapter presents the technologies and data foundation on which the methodology presented in Chapter 4 is based. The first section summarizes relevant software, and the second section describes the different data sources applied and their corresponding data sets.

3.1 Software and programming languages

All implementations and analyses behind the proposed methodology presented in Chapter 4 were conducted in Jupiter Notebook (Kluyver et al. 2016) with Python (Van Rossum and Drake 2009) as the main programming language. Jupyter Notebook is an open-source, web-based computing platform that enables combining software code with explanatory text and visualizations in the same document. Therefore, it is suitable for developing learning models where illustrative EDA combined with data implementations and model development are necessary. Python was chosen due to its ease of use and supportive online community. In addition, it comes with several frameworks and libraries that support scientific calculations and data analysis, of which the most important for this project were Pandas (McKinney 2010), Numpy (Harris et al. 2020) and Matplotlib (Hunter 2007) for data processing and data analysis and SciPy (Virtanen et al. 2020) for ML development.

3.2 Data foundation

The aim of the conducted data allocation was two-fold. Firstly, it was necessary to collect information describing vessel activity in the Arctic, which would serve as a basis for the prediction target in this thesis. Secondly, it was desired to identify explanatory predictors to Arctic activity patterns according to the research questions, including Arctic environmental

physics and climatic changes. An overview of all data sources investigated for this purpose is presented in Appendix A. The sources presented in this section were chosen among these based on the following criteria:

- The data are easily accessed
- The data have an Arctic coverage
- The data have similar spatial and temporal coverage as the data from the other sources already chosen
- The data are represented in a format feasible for merging with the data from the other sources already chosen

3.2.1 ASTD PAME

Vessel traffic data were collected from the ASTD system developed by PAME. The repository provides AIS records transmitted between 2013 and 2021 from IMO registered ships operating in the Arctic region, as marked in Figure 3.1. The ASTD AIS data were preferred over other activity data sources due to their strong Arctic focus and documented high-quality (*PAME - Arctic Ship Traffic Data* 2022). In addition, the motivation behind launching the system complies well with the objective of this thesis, as explained in Section 2.3.2.

Access to the ASTD data must be applied for and may be granted through three different levels. Level 1 gives access to the whole database, including vessel identity information by the IMO number and the Maritime Mobility Service Identity (MMSI) and detailed ship information. Level 2 and 3 access differ from Level 1 by excluding the ship identification numbers and having less detailed classifications of ship types. While Level 1 data are classified into over 200 ship types, Level 2 and 3 are aggregated to 50 and 13 ship types, respectively (PAME 2021d). NTNU was permitted the Level 3 access data containing the attributes given in Table 3.1. In addition to these attributes, the data include attributes of measurements on ship emissions and waste substances.

The AIS data are given in comma-separated value (CSV) files, representing the records in a tabular format. Each row corresponds to an AIS record, i.e., a positional signal given by a vessel, divided into columns representing the attributes in Table 3.1. The data are accessed through an FTP server which distributes individual monthly data sets.

Figure 3.1: Map of the ASTD area (red, dotted line). Note that the data coverage is not limited to the shadowed area, which is applicable for the Polar Code. Illustration retrieved from British Antarctic Survey (2020)

shipid	An anonymous, unique ship id to each month
date_time_utc	The time stamp of a signal
flagname	The flag of the ship
iceclass	The ice class based on the Finnish-Swedish Ice Class Rules (FSICR) (TRAFICOM 2019)
astd_cat	An aggregated category given by PAME (13 types in total)
sizegroup_gt	A size group by ground tonnage which yields the ship's volume or capacity and illustrates its size (7 groups in total)
fuelquality	The type of fuel used by the vessel (6 categories)
fuelcons	Fuel consumption estimation by an algorithm based on dynamical and ship specific information
dist_nextpoint	Distance to the next signal from the ship
sec_nextpoint	Seconds to the next signal from the ship
longitude	Positional longitude in WGS84 datum
latitude	Positional latitude in WGS84 datum

Table 3.1: List of attributes from the PAME AIS data

3.2.2 Copernicus

Copernicus is the European Union (EU) Earth observation program that offers open-source information provided by satellites and information collected from ground-based points of interest (*Copernicus* 2021). As indicated in Section 2.4, Arctic sea ice retreat is commonly associated with navigation access in the Arctic. As warmer temperatures and reduced sea ice are direct effects of the changing climate, information about Arctic sea surface temperature (SST) and sea ice concentration (SIC) was retrieved from the Copernicus Marine Service, the marine-focused component of Copernicus.

The Copernicus data are processed through four levels starting at the collection stage of raw satellite inputs to the final stage of an analyzed, interpolated product (Høyer and She 2007). Various pre-processing techniques and quality controls are conducted on the input data, such as only including cloud-free satellite data, subtracting sensor-specific biases, and replacing observations of SST with ice surface temperatures where the SIC is more than 70%. Additionally, the final level measurements have been validated against in-situ measurements from buoys and ship observations, showing a stable mean SST performance by standard deviations less than 0.7°C (Høyer, Riebergaard, et al. 2021).

The final analyzed data are spatially represented by geographical coordinates in the WGS84 projection with a 0.05 degrees resolution. The geographical and temporal coverages are north of 58°N and daily aggregations from January 1982 to May 2021, respectively. The products are accessed through Copernicus' FTP server and are given in a Network Common Data Form (NetCDF). NetCDF is an array-oriented representation commonly used for multidimensional geospatial data, where each file represents one daily aggregated observation (Høyer, Kolbe, et al. 2021). The NetCDF files retrieved from Copernicus contain the attributes given in Table 3.2. SST and SIC are denoted as *analysed_st* and *sea_ice_fraction*, respectively.

time	The time stamp of measurement
longitude	Positional longitude in WGS84 datum
latitude	Positional latitude in WGS84 datum
analysed_st	Analysed sea and ice surface temperature in Kelvin
sea_ice_fraction	Sea ice fraction

Table 3.2: List of attributes from environmental data, Copernicus

3.2.3 IBCAO

The underwater depth of the seafloor reflects safe vessel passages and may impact activity differently depending on the type of operation. For example, cargo ships may require a minimum depth for traversal, and fishing activities are attracted to areas of depth where certain species reside. Hence, geospatial information about the ocean floor was considered a potential explanatory activity predictor. Such data were retrieved from the International Bathymetric Chart of the Arctic Ocean (IBCAO) initiative, which has been the authoritative source of seafloor depth within the Arctic Ocean since 1997 (Jakobsson, L.A. Mayer, and Bringensparr 2020). Their initiative aims to develop a digital database and contribute toward increased understanding of tides, fishing resources, tsunami forecasting, ocean circulation, and environmental change. Their data are calculated through acoustic methods, where the time between sound pulses and their echos in response are measured from vessels. In addition, IBCAO uses satellite-derived measures in regions of sparse acoustic data coverage. This is done by observing anomalies in gravity which are correlated with the topography of the ocean seabed (*Gridded Bathymetry Data - Arctic Ocean (IBCAO)* 2020).

The data have a spatial coverage north of 64°N, and are represented as NetCDF files. As given in Table 3.3, the records are provided in an "xyz" format, where the "x" and "y" variables represent grid cell positions in Polar Stereographic projection coordinates, and the "z" variable corresponds to the elevation in meters of which the sea surface has a value of zero. The latest version of the data is available in regular grid sizes of either 200 meters or 400 meters (Jakobsson, L. Mayer, et al. 2020).

x	Horizontal position in Polar Stereographic projection
y	Vertical position in Polar Stereographic projection
z	Elevation in meters

Table 3.3: List of attributes from bathymetry data, IBCAO

3.2.4 NASA's OBPG

Distance measures implicitly explain the level of remoteness. As warmer temperatures arise in the Arctic, it is reasonable to assume that the subsequent increase in navigational access may cause maneuvers to venture further away from land. Like the bathymetric data, information regarding distances to the coastline may insinuate the type of industry. Whereas passenger

ships typically operate close to the coast for observations of species and nature, cargo ships traverse across deeper waters over longer passages.

The Ocean Biology Processing Group (OBPG) of the National Aeronautics and Space Administration (NASA) collects, processes, and distributes ocean-related products from satellite-based missions. In 2012, the group generated a global data set of distances to the nearest coastline using Generic Mapping Tools (GMT) software, an open-source collection of software packages for displaying and processing geospatial data (Wessel et al. 2019). The data have a grid resolution of 0.04 degrees and are available through a compressed text file. Three attributes represent the data records: longitude, latitude, and the distance measure, such that each record corresponds to a distance measured in kilometers from a specific spatial point (*NASA Ocean Biology Processing Group (OBPG)* 2012). The measurements provided by NASA will be denoted as the *distance data* in the remaining part of this thesis.

longitude	Positional longitude in WGS84 datum
latitude	Positional latitude in WGS84 datum
distance	Distance to the nearest coastline in km

Table 3.4: List of attributes from distance data, NASA OBPG

3.2.5 NMDC

The Norwegian Marine Data Centre (NMDC) is a marine research infrastructure that coordinates marine data from waters surrounding Norway. They aim to provide seamless access to historical marine data to contribute toward national and international quality research within marine science. The data center distributes open-source information describing both physical ocean characteristics and marine biology (Stenseth n.d.). It is expected that less sea ice in the Arctic will open up for plankton production, followed by new areas for fish stocks in the North (Hollowed, Planque, and Loeng 2013). Consequently, ecological changes could be linked to activity presence, especially within the fishing industry. Hence, it was desired to incorporate this ecological aspect into the data foundation of this thesis.

The NMDC provides fish catch data from the Barents Sea Ecosystem Survey, which monitors the status and changes in biological variables from the Barents Sea and adjacent waters. This is a Norwegian/Russian joint survey that has been run every autumn from 2004, of which data are available until 2019. The data cover the Barents Sea region and are aggregated by trawl

stations in a 35nm (nautical mile) grid resolution. The catch is manually measured onboard the trawl stations by count and weight (Johannesen et al. 2021). The resulting data files are provided by year in a tabular format where each row, or record, corresponds to measures from a specific trawl station given time and space. The first data columns, or attributes, represent physical and temporal information about the point of measurement, including latitude, longitude, time, station identification number, gear, depth, and tow distance. The remaining 84 columns represent the species of which cell entries are the catch measurements.

Chapter 4

Methodology

This chapter describes the methodology applied in this thesis which seeks to answer the first research question. The procedure is based on the technologies and the data foundation from Chapter 3 and structured according to the ML workflow explained in Section 2.6. In order to validate the applicability of the proposed solution and answer the remaining research questions, two case studies were defined from the processing steps and exposed to two tree-based ML models. The output from the models constitutes the experimental results of this thesis, which are further described in Chapter 5.

4.1 Definition of project domain

Activity as a measure of time and space is a broad term and needs a precise definition to qualify as an ML target. As expressed in Section 2.4 and 2.7, there are multiple ways of measuring the volume of vessel traffic. Hence, the ML problem can be formulated either as a regression or a classification problem. Potential activity measures for regression problems could be values directly regarding ship presence, such as vessel density, vessel frequency, and operational time spent, or measures coming as a *result* of ship presence, such as fuel consumption, vessel speed, and emissions. On the other hand, activity prediction by classification could be considered as, for example, the binary task of classifying activity presence or absence, vessel type, or size group.

4.1.1 Prediction target

To answer the second and third research questions, the target was identified through two case studies representing different activity perspectives. The first case study addresses the

binary task of predicting vessel presence within the chosen period and area. The second case study considers the spatial locations of vessel presence only, and targets the vessel density by *unique ship count* from the aspects of three different industries: fishing, cargo shipping, and tourism. As such, the two case studies, or ML tasks, satisfy the fact of being entirely dependent on *external* climatic changes and not vessel characteristic information provided by AIS transmitters. In addition, a target definition based on vessel density contributes toward understanding the changing Arctic risk picture, as discussed in Section 2.2. Due to discontinuities in the distribution of unique ship count, which is further elaborated in Section 4.2.6, the target of Case Study 2 was binned into a classification problem considering three degrees of vessel density: high, medium, and low. The tasks of each case study are explicitly formulated in Table 4.1. The I notation conveys that the temporal and geographical boundaries are dependent on the given industry. While G considers the whole geographical boundary as explained in Section 4.1.3, G_I is restricted to only considering the respective industry's spatial locations of vessel presence.

Case study 1	**Case study 2**
Given time intervals T and spatial grid cells G, predict class y_{1ab} fore each t_a, g_b where $y_{1ab} \in \{presence, absence\}$	Given time intervals T_I, spatial grid cells G_I and industries I, predict class y_{2abc} for each t_{Ia}, g_{Ib}, i_c where $y_{2abc} \in \{low, medium, high\}$

Table 4.1: Classification task formulations

4.1.2 Temporal boundaries

The temporal and geographical boundaries, corresponding to T and G, respectively, were defined in order to put the classification tasks into context. This thesis covers a time region from January 1st, 2015, to June 1st, 2021, except for the fishing industry in Case Study 2, where the upper boundary is January 1st, 2020. The boundaries were selected according to the temporal coverage of the data records collected. Although the ASTD AIS data cover vessel traffic from 2013 until the present, this thesis is focused on the records from 2015 and above due to fewer AIS ground stations and satellites in the early years as well as the new AIS requirements taking place from 2014, as stated in Section 2.3.3. The upper time limit was restricted by the SST and SIC data from Copernicus. However, the temporal coverage of the NMDC fish catch data further restricted the upper boundary for fishing vessel density prediction. Hence, $T_{fishing}$ differs from the rest.

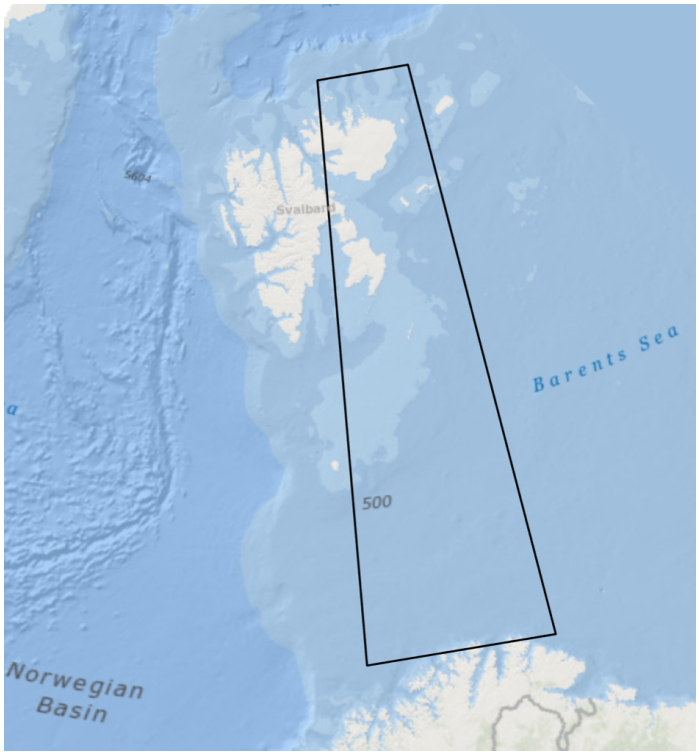


Figure 4.1: Spatial boundaries of the target area. Restricted by longitudes 20° and 30° and latitudes 71° and 81°

4.1.3 Geographical boundaries

The spatial area selected is a region in the Barents Sea extending from the northern coast of Norway to above the Svalbard islands, more specifically defined by the longitudes 20° and 30° and latitudes 71° and 81° as seen in Figure 4.1. In order to discover environmental and physical impacts on vessel presence and density, it was desired to choose a spatial area affected by such variations. The Barents Sea was considered a suitable area according to the climatic changes caused by the polar front and the various industries operating in the region. The fishing and tourism industries are well represented due to high fish productivity and cruise offers to Svalbard to the northwest and Franz Josef Land to the northeast. Moreover, the Barents Sea is exposed to cargo shipping activities as the NSR passes through the region.

Due to time processing issues and computer memory errors caused by the vast amounts of data from the whole Barents Sea region, there was a need for further spatial reduction. To include the climatic variations separated by the polar front, tourism associated with Svalbard and Franz Josef, as well as shipping activities related to the passage close to the Norwegian coast,

it was decided to perform a longitudinal reduction, resulting in the final spatial boundaries as given.

4.2 Data processing

The methodology in this thesis relies on real-world information from multiple sources, measured by sensors and satellites. Geographical and temporal gaps and erroneous samples most certainly hide within such large volumes of data. Hence, a thorough EDA was conducted in order to gain insight into the data and discover potential irregularities. According to this analysis, the data were subject to several processing steps to increase prediction capabilities, reduce unnecessary information and facilitate data management and applicability. The following section explains the actions applied to obtain the proposed training data set, which are justified by a selection of relevant plots from the EDA. Additional outcomes from the analysis are further attached in Appendix B.

4.2.1 Data allocation and size reduction

As elaborated in Chapter 3, the data are provided in different data formats from, in total, five sources. Monthly AIS streams, SST, and SIC data were downloaded individually from PAME's and Copernicus' FTP servers. Yearly fish catch data, distance data, and bathymetry data by a 400 meters grid resolution were gathered directly from their respective web pages. The IBCAO 400-meter grid spacing has a favorable file size and the highest resolution among all data sources. Hence, it was chosen over the even higher 200-meter grid spacing alternative. Both SST, SIC, and bathymetry data are provided in NetCDF formats and were converted by the *netCDF4* (2015) library to CSV files to coincide with the other data sources. The CSV format was preferred due to its compatibility with the Python libraries for data analysis and ML modeling. Further, the data files from each source were restricted to only cover measures within the spatial boundaries.

As recommended by PAME (2021d), all vessel signals with less than ten positions per month were filtered out from each data file. This ensured exclusion of signals most likely transmitted from outside the Arctic area yet randomly picked up by satellites. The records from outside the spatial boundaries were removed from each monthly data file before the data were concatenated to a final CSV file covering all months within the selected period from Section 4.1.2. As a result, the initial file size of all AIS samples was reduced from 263 GB to 7 GB, and a final CSV file of approximately 29 million records.

Thesis category	ASTD category
Fishing vessels	Fishing vessels
Passenger ships	Passenger ships
	Cruise ships
Cargo ships	General cargo ships
	Refrigerated cargo ships
	Ro-Ro cargo ships
	Container ships
	Bulk carriers
	Gas tankers
	Chemical tankers
	Oil product tankers
	Crude oil tankers

Table 4.2: Aggregation of ASTD categories

As stated in Chapter 3, the ASTD Level 3 data distinguish ship types according to 13 ASTD categories. However, Case Study 2 focuses on three main vessel types, namely fishing vessels, cargo ships, and passenger ships. Hence, a further aggregation of the 13 ship types was conducted, as presented in Table 4.2. The division was based on the overall objectives of the three main ship types, where all ship types within cargo ships were considered as related to sea freight. The final "Other activities" ASTD category was not included within any of the main categories as it covers over 20 different vessel types of different behavior, including mooring buoys, yachts, sailing vessels, research vessels, and others.

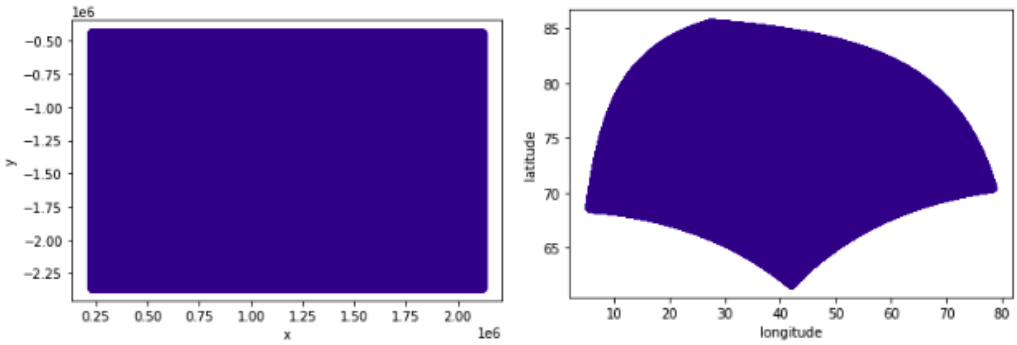
4.2.2 Gridding

The individual data sources provide different information corresponding to features important for target prediction. Whereas all sources include a pair of two-dimensional spatial coordinates, the AIS streams, the fish catch data, and the SIC and SST measures have an additional temporal dimension. These three dimensions constituted the common key variables for joining the multiple sources into one composite data set. However, as stated in Chapter 3, the spatial and temporal values have different resolutions and granularity, respectively. Hence, it was necessary to assign the records into grids to enable data merging from the different sources.

Gridding is the procedure of interpolating data samples onto a regular grid, in this case by three dimensions, as both space and time are considered. Each grid cell, or data pixel, corresponds to a value associated with a time frame and region on the surface of earth (Davis and Sampson 1986). When several data samples fall within the same grid cell, a mathematical function will calculate the associated grid cell value, such as the mean, the maximum, or the minimum of the belonging values. Hence, the level of detail represented by the grids is dependent on the cell size, i.e., the spatial and temporal resolution.

In order to calculate vessel densities by unique ship count, it was necessary to separate the vessels from each other. However, the ASTD Level 3 data does not provide the identifying MMSI numbers but an alternative ship-id unique to each month. This means that each ship-id within a specific month corresponds to one vessel only, but AIS transmissions between months may share the same id although they originate from different vessels. As such, the temporal resolution was defined as a *monthly* time-frequency according to the ship-ids validity. The choice of spatial resolution, on the other hand, was determined by a trade-off between being small enough to capture necessary data details but large enough to prevent issues related to computer storage and execution time. A trial-and-error approach was conducted by testing several spatial cell sizes, of which a 15kmx15km resolution yielded the most acceptable precision and time complexity.

The regular spatial grid was created by dividing the area restricted by the geographical boundaries into equal squares of the predefined cell size. Each pixel was assigned a unique grid index and an individual pair of coordinates, represented by the coordinates of the cell's center. The bathymetry data were used to neglect grid cells located on land, as all bathymetry values below or equal to zero correspond to spatial locations at sea. As such, land-based measures and potential erroneous AIS messages from land could be tracked by the data points not belonging to any grid cell. However, in contrast to the other data sources, the original bathymetry data are given in meters by the Polar Stereographic projection. Hence, the data were transformed to the WGS84 projection using *PyProj* (2022), as illustrated in Figure 4.2. Then, all bathymetry values above zero were rejected and used to remove the corresponding land-based grid cells. The resulting spatial grid data were stored as a two-dimensional DataFrame, which is a tabular data structure supported by the Pandas library (McKinney 2010), with a total of 21486 spatial grid cells located at sea. When paired with the temporal dimension of months within the chosen period, the total number of grids in three dimensions became 1654422.



(a) Meters by Polar Stereographic projection (b) WGS84 coordinate system with reference code EPSG4326

Figure 4.2: Before and after geographical transformation of IBCAO data. The transformation was performed before further spatial restriction by the selected boundaries within the Barents Sea

4.2.3 Irrelevance and noise

The created regular grid was used to map the samples from the data sets to their belonging grid index. As such, AIS messages, SST and SIC measures, and distance data on land were removed by extracting the samples not associated with any grid index. This procedure is illustrated by the distance data in Figure 4.3 where the samples located at the Svalbard island, in the upper left corner, were removed. However, the islands at latitude 28° and longitude 77° were not removed because they were too small according to the created grid resolution, meaning that the grid cells covering the islands involve bathymetry values both above and below the sea surface.

746700 duplicates were discovered among the AIS records. As further elaborated in Section 2.6.1, the duplicates would provide no additional valuable information for target prediction and were therefore removed. In addition, it was desired to investigate whether there were cases of different vessels within the same month having similar ids. This was done by extracting a subset from the AIS data, which included the static features only, i.e., flagname, ice class, category, and size group, and inspecting whether similar ids within a month differed by these static features. However, no such instances were identified, which justifies the ASTD data's documented quality.

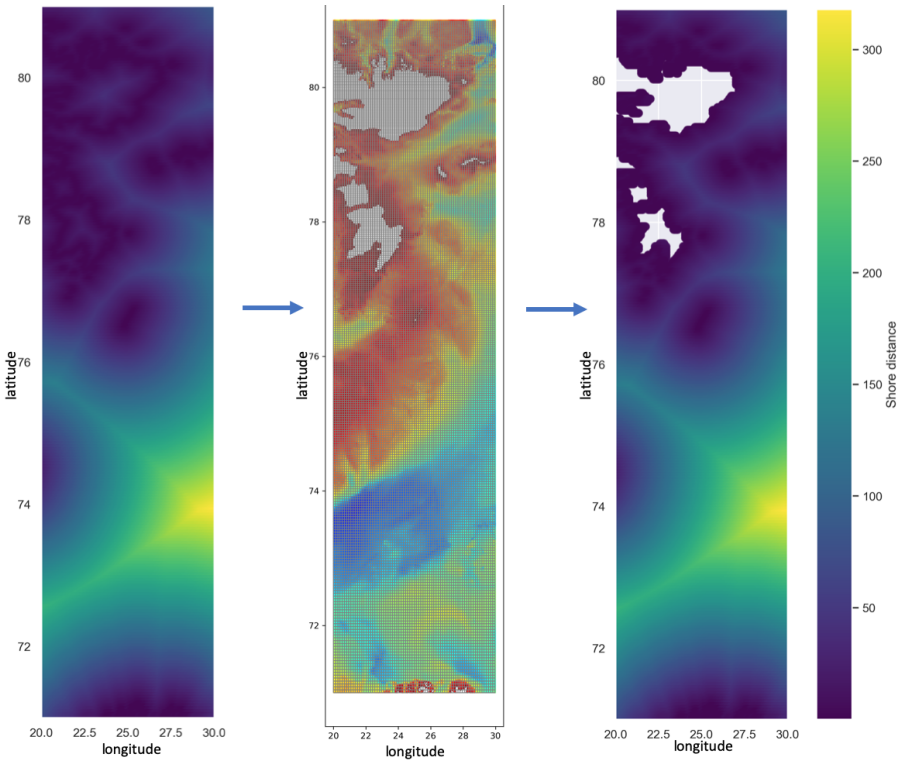
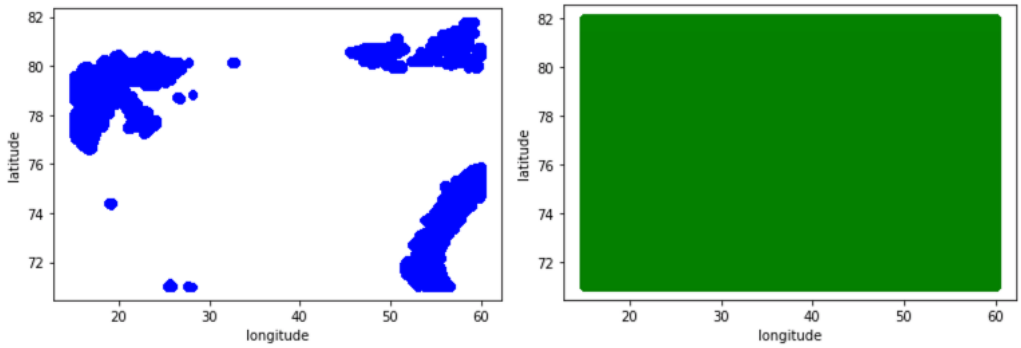


Figure 4.3: Removal of land-based positions from distance data by grid mapping. The background map of the grid in the middle corresponds to the ocean depths which were used for removal of grid cells positioned on land

4.2.4 Missing values

Both the SST and SIC data involve several NaN values, as illustrated by space in Figure 4.4 which covers the entire Barents Sea region. Whereas the plots show that the clusters of non-recorded sea ice values are restricted to land, which is reasonable, the sea surface temperatures are spread over the whole region. Therefore, it was assumed that the missing SIC values are MNAR while the missing SST are MCAR. Consequently, many MNAR values were neglected when the data samples were mapped to the spatial grids, and land-based records were removed. However, there remained temporal gaps by dates in both measures. The 21486 grid cells combined with 2342 dates within the selected months from 2015 to 2021 constitute $2342 \cdot 21486 = 50320212$ three-dimensional combinations. By comparing these combinations with the corresponding combinations of *existing* measures in the data, it was detected that a total of 19407615 SIC values and 20038560 SST values were missing. The missing values were handled by spatial and temporal interpolation, represented by grid cells and dates, respectively, where a linear relationship was assumed for small changes in time



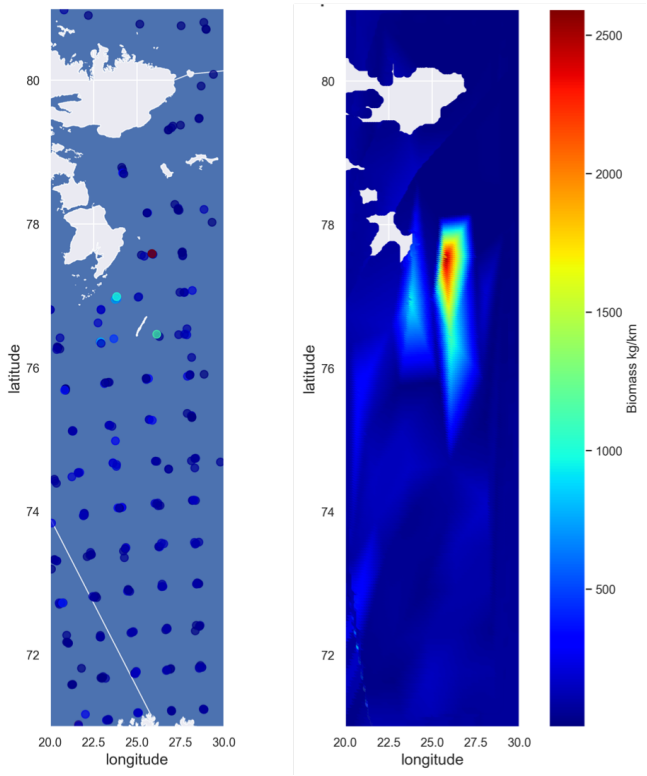
(a) Missing SIC values by space (b) Missing SST values by space

Figure 4.4: Spatial plots of missing sea surface temperature and sea ice measures. NaN values within the whole time frame are included, which, altogether, cover the entire area by SST

and space. A multidimensional piecewise interpolation package from Virtanen et al. (2020) was applied, which creates a reference object, i.e., a convex hull, by the existing values for triangulation of the missing values. However, some of the requested interpolation points remained outside the three-dimensional convex hull and were handled by the nearest neighbor technique. The nearest neighbor returns the closest value to the point of interpolation in time and space, hence was considered a suitable approach as temperature and ice extent are expected to have minor local variations.

The NMDC fish catch data have a coarser distribution both spatially and temporally compared to the other data sources. This is because the fish catches are measured manually instead of continuously tracked by satellites or sensors. As illustrated by Figure 4.5a, the fish catch data resolution of 35nm corresponds to 64.82km. Hence, multiple cells of missing fish catch measures were created when the data were assigned to the spatial grids. However, as the fish catch resolution corresponds to the distributions of trawl stations, it is reasonable to assume that the missing cell values do not refer to a lack of catch. Instead, they correspond to the sources of the final fish catches measured by their nearest station. Consequently, the missing cell values were handled by interpolation by the same approach described above, resulting in the spatial catch distribution as given by Figure 4.5b.

Figure 4.6a shows that the ASTD data suffer from high ratios of missing vessel characteristic values, especially within the *iceclass* attribute. However, according to the objective of this thesis and the target values defined as such, the interior features of the ASTD data were not considered meaningful predictors. This is because they are provided along with the vessel's AIS streams, hence are unknown unless the vessel and its location are identified. Consequently,



(a) Original spatial distribution. The blue background area corresponds to the sea

(b) Spatial distribution after interpolation

Figure 4.5: Before and after interpolation of the NMDC fish catcg data

their ratios of missing values were irrelevant for the final composition of the ML data set. On the other hand, it was necessary to track any discrepancies in the *astd_cat* attribute, as the training data sets of Case Study 2 are based on the ASTD Level 3 category. In fact, it was identified only one single vessel categorized by NaN. The AIS signals transmitted from this vessel were plotted and resembled a typical cargo ship traversal through the NSR. The plot is not presented in this report according to vessel identity confidentiality. In addition, the vessel's value of *sizegroup_gt* yields "10000-24900 GT", which is only represented among the other labels categorized as cargo ships except for a few cruise ships. Hence, it is reasonable to assume that the NaN categorized vessel corresponds to a type of cargo ship, and the value was thereby manually replaced by this category.

(a) All AIS records

(b) Records within "Unknown" category

Figure 4.6: Missing value ratios of static ASTD values

In addition to the NaN-categorized vessel, it was discovered from the EDA that 5772041 AIS records, transmitted from 4729 unique vessels, are categorized as "Unknown" by the PAME group. Figure 4.6b also proves that most of the static features describing vessel characteristics within the "Unknown" category are missing. In fact, all missing values of *sizegroup_gt* within the AIS streams are related to this category. Moreover, Figure 4.7 illustrates that there is no apparent connection between the overall movement patterns of the "Unknown" vessels. This observation substantiates the assumption that the "Unknown" category is randomly missing, i.e., it does not belong to one specific type of vessel. However, classifying the "Unknown" labels to an industry based on their ship characteristic features would have caused erroneous assumptions and induced a biased model due to their respective high ratios of missing values. One potential solution involves applying more advanced ML approaches, such as deep learning, to distinguish the vessels by their spatial behavior. However, such applications go beyond the scope of this thesis. Consequently, the "Unknown" vessels were excluded entirely from Case Study 2 but remained part of the training data for Case Study 1 as this classification task is irrespective of type.

4.2.5 Feature transformations

Similar length units: Values of bathymetry, distance data, and towing distance are represented by length units. However, these metrics are expressed by meters, kilometers, and nautical miles, respectively. In order for the ML models to better understand the relationship between the distances, the features were transformed into kilometer distances. In addition, the bathymetry values were converted to positive values to better interpret potential correlations with the prediction target. Similarly, the SST values were transformed from Kelvin to Celsius.

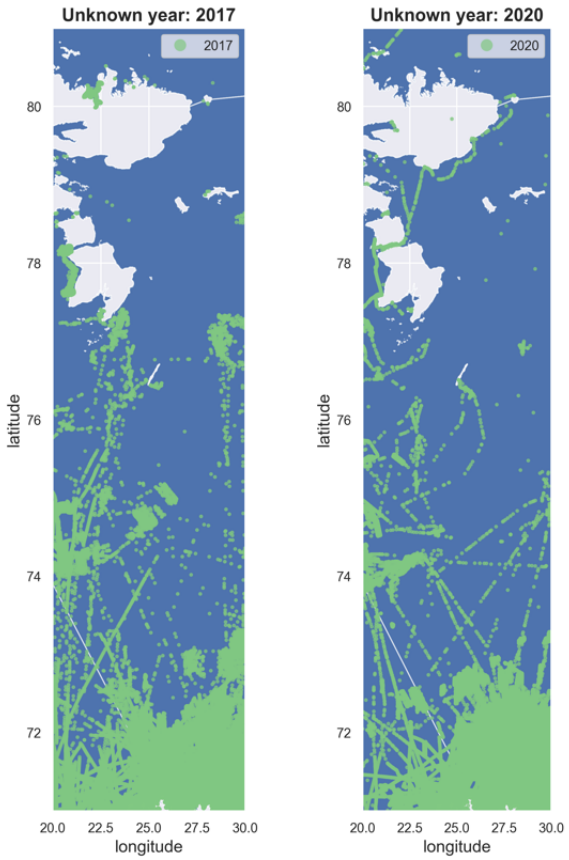


Figure 4.7: Trajectories from "Unknown" categorized vessels year 2017 (left) and 2020 (right). The years are randomly selected for illustration. The blue background area corresponds to the sea

Biomass per distance: The fish catch data cover biomass from 84 different species, represented by the individual columns. However, the biomass per specie varies significantly according to the specie's rarity. In addition, having many input features increases the data dimensionality, causing the predictive modeling task to become more challenging (Goodfellow, Bengio, and Courville 2016). Ultimately, it is reasonable to assume that only the commercial species constitute meaningful predictors for fishing activities. Based on these considerations, it was determined to subset the fish catch data to commercial species only and extract a new feature as the aggregated sum of the total commercial biomass. Table 4.3 presents the commercial species identified in the data according to ICES (2022) and Shevelev and Gjøsæter (1999). The catches of the corresponding subset data were standardized by their respective towing distance in kilometers and aggregated to a final *biomass* feature in kilograms per kilometer towed.

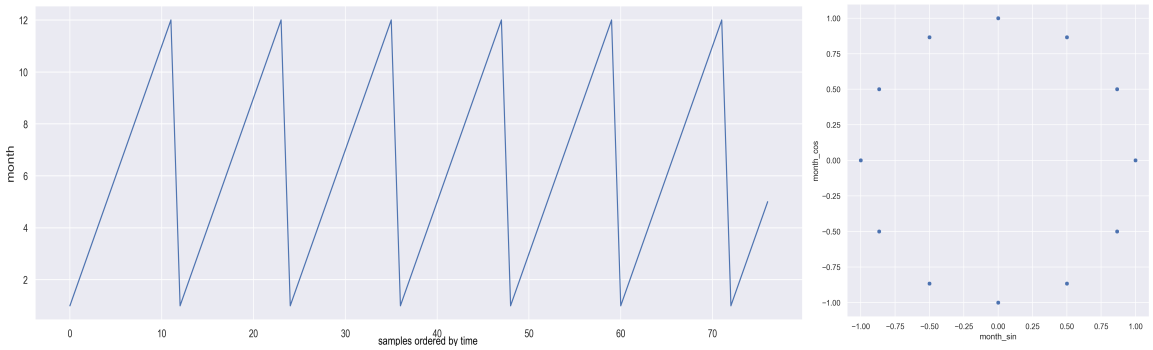
Common name	**Scientific name**
Arctic skate	Amblyraja hyperborea
Arctic cod	Boreogadus saida
Atlantic wolffish	Anarhichas lupus
Beaked redfish	Sebastes mentella
Capelin	Mallotus villosus
Cod	Gadus morhua
Golden redfish	Sebastes norvegicus
Greenland halibut	Reinhardtius hippoglossoides
Haddock	Melanogrammus aeglefinus
Herring	Clupea harengus
Leopardfish	Anarhichas minor
Long rough dab	Hippoglossoides platessoides
Lumpfish	Cyclopterus lumpus
Northern wolffish	Anarhichas denticulatus
Norway redfish	Sebastes viviparus
Saithe	Pollachius virens
Spinytail skate	Bathyraja spinicauda
Thorny skate (starry ray)	Amblyraja radiata

Table 4.3: Commercial species extracted from NMDC data

Cyclic time dimension: The temporal dimension of the data is represented by categorical time encoding through the *year* and *month* attributes. Whereas the years constantly increase from 2015 to 2021, the months occur in cycles. However, this natural cyclic time behavior is not apparent to the ML model. When the months are encoded as integers ranging from 1 (January) to 12 (December), there are discontinuities in the feature distribution at year-end due to the sudden reset point, visualized in Figure 4.8a. Hence, the cyclic behavior was encoded explicitly so that the model could understand the connection between January and December.

$$
\begin{aligned}
x_{sin} &= \sin\left(\frac{2\pi x}{max(x)}\right) \\
x_{cos} &= \cos\left(\frac{2\pi x}{max(x)}\right)
\end{aligned}
\tag{4.1}
$$

As expressed in 4.1, the cyclic encoding was performed by transforming the months, $x \in [1, 12]$, into two dimensions using the sine and cosine periodic functions. Both dimensions were necessary, as the periodic symmetry from only one dimension would cause two records to have the same transformed value. Consequently, the cyclic month pattern presented in Figure 4.8b was obtained.



(a) Original evolution of month values with jump discontinuities

(b) Month values in two dimensions

Figure 4.8: Cyclic transformation of *month*

4.2.6 Data aggregation and target extraction

After the individual data files were processed and mapped to grids, they were ready to be combined into complete ML training sets. The common key variables used for data merging were *year*, *month* and *grid_index*. Since the original data files are represented by different time and space resolutions, it was necessary to down-sample the data by the key variables to enable proper merging. In addition, the classification targets presented in Section 4.1.1 were not expressed explicitly by the AIS streams and had to be quantified through data aggregation. Hence, the data from each source were grouped by the *grid_index* attribute and the temporal attributes if these existed. Then, the numerical features, i.e., bathymetry, distance, SST, SIC, and biomass, were aggregated by the mean values of their respective samples associated with a given combination of key variables. On the other hand, vessel density, denoted as *unique ship count*, was extracted from the AIS streams by counting the number of unique vessels operating within each *year*, *month* and *grid_index*. The final number of aggregated records according to the key attributes represented in each data set are presented in Table 4.4. Since the upper boundary of the period is May 2021, the last five months are added separately for the calculations on the Copernicus data sets. On the other side, the number of total records within the AIS data is not similarly calculated by the number of grids and months since vessel activity is neither present in all 21486 grid cells nor in all possible combinations of time and space. This also explains why the number of unique *grid_index* values in the AIS data differs from the other sources.

Data source (feature)	**#years**	**#months**	**#grid cells**	**#total records**
IBCAO (bathymetry)	None	None	21486	**21486**
NASA (distance)	None	None	21486	**21486**
NMDC (biomass)	5	None	21486	$5 \cdot 21486$ $= \mathbf{107430}$
Copernicus (SIC & SST)	7	12	21486	$(6 \cdot 12 + 1 \cdot 5) \cdot 21486$ $= \mathbf{1654422}$
ASTD (unique ship count)	7	12	20030	**463162**

Table 4.4: Number of records ("#") within each data set as a result of unique years, months and grid cells

Before the individual data sets were merged, the AIS streams of all vessel types were subdivided into four different representations according to the case studies: whereas all processed AIS records remained for Case Study 1, three additional subsets were extracted according to the fishing, cargo shipping, and tourism industries in Case Study 2. Then, each of the four AIS data sets was separately combined with the other data sources as follows:

Case study 1

1. SST and SIC data were merged together by *year*, *month* and *grid_index*.
2. ASTD data were merged *into* the composite SST and SIC data by *year*, *month* and *grid_index*.
3. Bathymetry and distance data were individually merged *into* the composite SST, SIC and ASTD data by *grid_index*.

Case study 2

1. SST and SIC data were individually merged *into* the ASTD data by *year*, *month* and *grid_index*.
2. Bathymetry and distance data were individually merged *into* the composite ASTD, SST and SIC data by *grid_index*.
3. **Fishing vessels only**: biomass data were merged *into* the composite ASTD, SST, SIC, bathymetry and distance data by *year* and *grid_index*.

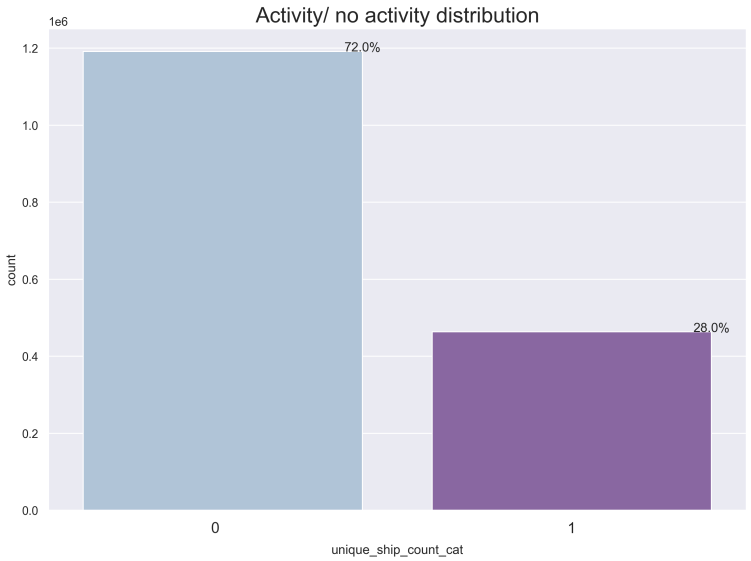
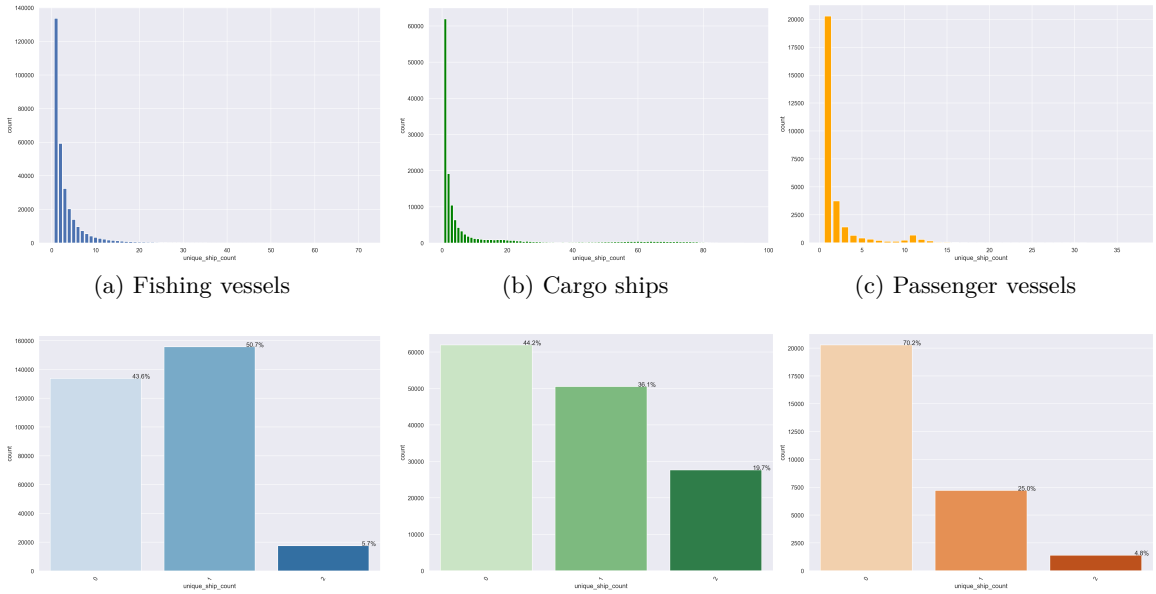


Figure 4.9: Target distribution of Case study 1. The percentages correspond to the respective class' proportion of records within the data set

Since the SST and SIC data cover the entire spatial region and time, these were first combined according to Case Study 1. When the AIS records were joined *into* this data set, several combinations of values of month and *grid_index* remained without any unique ship count value, i.e., NaNs, meaning activity *absence*. Correspondingly, the samples that resulted with existing unique ship count values were related to activity *presence*. By replacing the NaNs and unique ship count values with 0 and 1, respectively, the target values of Case Study 1 were obtained, yielding the distribution as given in Figure 4.9. In total, the data set for Case Study 1 contains 16554422 instances, of which 463162 are presence values. According to Table 4.4, the former corresponds to the total number of values in the spatiotemporal grid, while the latter corresponds to the aggregated AIS records.

In Case study 2, on the other hand, the aim is to target the amount of activity given activity presence and type of industry. Since the AIS records cover all spatial and temporal combinations of activity presence, the other data sources were merged *into* the AIS data. Hence, the unique ship count value has a minimum of at least one unique vessel. The upper level of Figure and 4.10 shows the number of data records of each unique ship count value for Case Study 2. The plots are color-coded according to industry. The data set of fishing vessels contains the most instances of 391567 records, followed by cargo ships and passenger ships of 140107 and 28881 records each, respectively. The distributions evolve rather similarly and contain discrete integers that are lower bounded by one vessel by month and grid cell and



(a) Fishing vessels (b) Cargo ships (c) Passenger vessels

(d) Fishing vessels (e) Cargo ships (f) Passenger vessels

Figure 4.10: Original distributions of unique ship count (a, b, c) and binned classes of unique ship count (d, e, f). The percentages correspond to the respective class' proportion of records within the data set

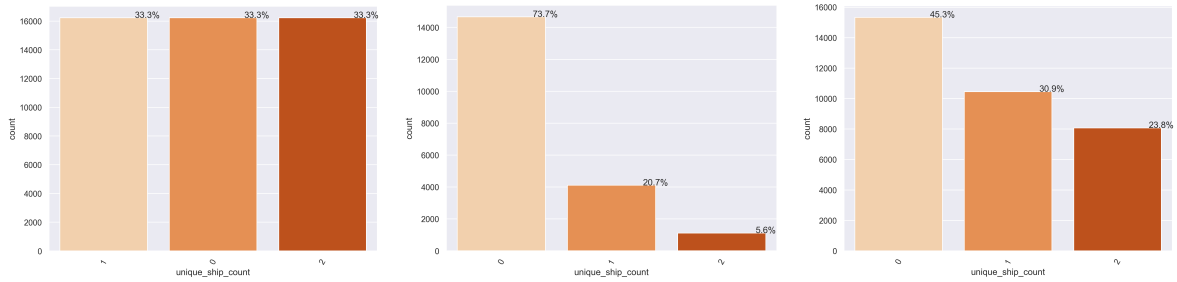
upper bounded by 37, 71, and 95 unique ships for passenger ships, fishing vessels, and cargo ships, respectively. However, each distribution involves several gaps within the value range. Hence, the initial target distributions may confuse any regression model due to expectations of predicting continuous quantities within an infinite set of values. Consequently, it was decided to transform the unique ship count values into categorical bins of three target classes representing different levels of activity density. This would create a more confident prediction and remove noise caused by discontinuities in the data.

The bin edges were defined based on the distributions seen in the upper level of Figure 4.10 and are presented below. Due to the significant dominance of one unique vessel by month and grid cell, followed by a rapid decrease toward ten unique vessels, as well as rare occurrences of unique ship counts above 30 within all industries, the division was conducted as such. The classes in ascending order refer to low, medium, and high levels of activity, respectively. The lower level of Figure 4.10 shows the resulting binned distributions.

- **Class 0** (low) corresponds to records where $unique\_ship\_count = 1$
- **Class 1** (medium) corresponds to records where $unique\_ship\_count \in [2, 10]$
- **Class 2** (high) corresponds to records where $unique\_ship\_count > 10$

4.2.7 Data imbalance

Real-world data commonly exhibit uneven class distributions, and the AIS streams used in this thesis are no exception. The ship movements differ in space and time, causing the class imbalances as visualized in Figure 4.9 and 4.10 for Case Study 1 and 2, respectively. Whereas Case Study 1 and cargo ships within Case Study 2 have fairly mild degrees of imbalance, fishing vessels and passenger ships are significantly affected by a 5% and 6% proportion of their respective minority classes (Class 2). As described in Section 2.8.2, it is preferable to resolve the imbalance problem prior to ML modeling as an uneven target distribution may confuse the model toward erroneous predictions of the majority class. Hence, improved ratios of the target classes were created using the SMOTETomek sampling method provided by the Python library Imblearn (Lemaître, Nogueira, and Aridas 2017). The SMOTETomek technique was selected based on the trade-off between achieving an evenly distributed target while neither overexposing the data to synthetically generated samples nor removing excessive amounts of valuable information.



(a) Oversampling by SMOTE (111% increase in samples)

(b) Undersampling by TomekLinks (14% decrease in samples)

(c) Hybrid sampling by SMOTE-Tomek (47% increase in samples)

Figure 4.11: Testing of resampling methods on passenger vessel data. The percentages correspond to the respective class' proportion of records within the data set

The data were first exposed for oversampling by SMOTE, which provided even target distributions, yet, significant data size increases. This is illustrated by Figure 4.11a, which shows a 111% increase in passenger vessel records, being the most unevenly distributed data set. On the other hand, undersampling by TomekLinks reduced the data by 14%. However, it did not significantly reduce the majority class compared to the minority classes. Finally, the hybrid method SMOTETomek proved the most favorable distribution. Figure 4.11c shows the passenger ship records resampled by SMOTETomek. The increase of records by 47% mainly affected the minority class, yet, the natural relationship between the classes was maintained.

4.2.8 Overview of data pre-processing

The steps described in this Chapter for developing the ML prediction data sets for each case study are summarized in Figure 4.12. All four data sets are based on the same predictors, but the prediction targets differ. The fishing vessels data include *biomass* as an additional predictor, which limits the upper time boundary to January 1st 2020. The composition of predictors and target values used for the respective ML tasks are provided in Table 4.5.

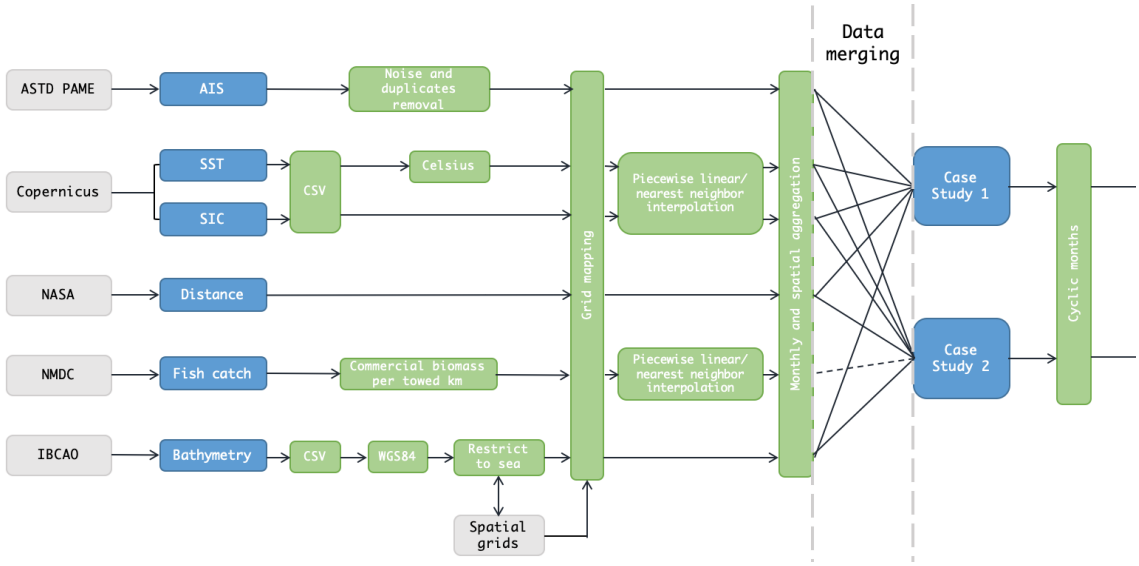


Figure 4.12: Flowchart of the data pre-processing phase. The stippled line in the merging step refers to the inclusion of NMDC data for the fishing industry in Case Study 2

4.3 Model development

After the data were investigated, processed, and combined, the final composite data sets were ready for individual model learning. This section explains how the second central part of the ML workflow, namely the ML model development, was performed in this thesis. This includes selecting suitable learning algorithms, exposing the algorithms for training data, testing their learning capabilities, and evaluating them. Similar approaches were applied for both case studies; hence, the notations "data" or "data sets" refer to all data irrespective of case study, unless stated otherwise.

Predictor attribute	Type	Description
year	int	Year within the time period of 2015 to 2021 (upper bounded by 2019 for fishing vessels)
month_sin	float	Sine transformed month value $\in [-1, 1]$
month_cos	float	Cosine transformed month value $\in [-1, 1]$
latitude	float	Positional latitude coordinate of a grid cell's center (WGS84)
longitude	float	Positional longitude coordinate of a grid cell's center (WGS84)
analysed_st	float	SST in degrees Celsius
sea_ice_fraction	float	SIC in percentage $\in [0, 1]$
bathymetry	float	Depth of the ocean floors in km, lower bounded by 0
dist_to_shore	float	Distance to nearest coastline in km, lower bounded by 0
biomass (fishing vessels only)	float	Biomass of commercial species in kg/towed km
Target attribute		
Case study 1 *unique_ship_count_cat*	int	Class 0 and 1 yield absence and presence of activity, respectively
Case study 2 *unique_ship_count*	int	Class 0, 1 and 2 yield low, medium and high levels of activity, respectively

Table 4.5: Final composition of predictors and target for ML modelling

4.3.1 Model selection

Reputations and previous documented experiences from the literature where ML problems of similar predictors are investigated were used as a starting point for selecting suitable supervised models for the given classification tasks. As presented in Section 2.7, many similar approaches are being applied in various contexts of AIS-based ML prediction. These include ANN, SVM, Decision Tree, Extreme Gradient Boosting (XGBoost), and RF, which additionally have proven to be among the most commonly used within several domains (Cracknell and Reading 2014; M. Liu, M. Wang, et al. 2013). XGBoost is a Gradient Boosted algorithm developed by Chen and Guestrin (2016), which scales well to extensive data as it creates branches in parallel while constructing each sequential tree. Compared to general Gradient Boosted methods, it uses more advanced computations for optimizing the model, such as the second-order gradients of the sequential residual for error minimization. In addition, the algorithm includes a regularization term which increases model simplicity and prevents overfitting (Sarker 2021). In order to detect which of the algorithms to use, a systematic comparison was performed, which is summarized in Table 4.6.

ML model	Strengths	Weaknesses
ANN	• Works well with large data sets (many records) • Supports complex, untructered non-linear data	• Complex to interpret • Sensitive to hyperparameters • Computationally demanding • Learns slowly
SVM	• Works well with high-dimensional data (many features) • Supports non-linear data	• Sensitive to hyperparameters • Inefficient on large data sets • Memory intensive
XGBoost	• Works well with large data sets (many records) • Robust to overfitting • Robust to irrelevant features	• Sensitive to hyperparameters • Sensitive to outliers
RF	• Works well with high-dimensional data (many features) • Robust to overfitting • Robust to noise and outliers • Stable against hyperparameters	• Complex to interpret • Computationally demanding • Learns slowly
Decision Tree	• Works well with large data sets (many records) • Highly interpretable	• Prone to overfitting • Learns slowly • Sensitive to changes in the data

Table 4.6: Comparison of ML models by strengths and weaknesses

The data sets for model training consist of several hundred thousand records but have a low-dimensional feature space. In addition, the targets suffer the imbalance problem, being skewed towards the lower classes. Based on these considerations and the outcomes from Table 4.6, the SVM and Decision Tree algorithms were rejected as the former handles dimensionality better than large data sizes, and the latter may easily overfit and prefer the majority class. RF and XGBoost were considered better candidates as these improve the robustness of one decision tree by generating multiple trees. ANN was also evaluated as a good candidate based on its ability to handle large data files of complex and potentially noisy data (Mitchell 1997). Conversely, ANNs require feature scaling and high processing power; hence, the ensemble methods were considered more feasible according to available data and hardware sources. Moreover, XGBoost has proven to outperform ANN in multiple winning solutions of data mining competitions (Chen and Guestrin 2016).

In the end, it was decided to explore both ensemble methods, RF and XGBoost, for further model training. A major benefit of the tree-based models is that they work well with uneven target distributions and non-linear data. Both models were selected as they share the fundamental concept of generating multiple decision trees for prediction. However, their strengths and weaknesses may offset each other (Sarker 2021). Whereas RF is stable to changes in hyperparameters, easy to tune, and difficult to overfit, the performance of XGBoost may fluctuate according to the choice of hyperparameters, and is more sensitive to overfitting.

On the other hand, XGBoost has the advantage of detecting and handling prediction errors during learning and often stands out in model performance (Chen and Guestrin 2016).

4.3.2 Model training and hyperparameter optimization

The ensemble classifiers were built by using the estimators *RandomForestClassifier* from the Scikit-Learn library (Pedregosa et al. 2011) and *XGBoostClassifier* from the XGBoost library (Chen and Guestrin 2016) which is compatible with Scikit-learn Application Programming Interface (API). First, the pre-processed data were split into a test set and a training set by a 20%/80% ratio, respectively. The split was performed by a stratified approach, which preserves a similar proportion of samples within each class as the original data set. Hence, a representative test set for evaluation was obtained according to the class distribution of the training set.

As explained in section 2.5 the test set yields an unbiased evaluation of model performance and should therefore not be exposed to any modifications after the split is done. In order to obtain resampling by SMOTETomek on the training data only, a pipeline for each classifier was constructed, which applies the resampling transformation and the learning process to the training data in sequence. Before individually tuning the classifiers' hyperparameters, each pipeline was trained and validated by cross-validation and compared against the validation of a corresponding pipeline without the resampling step. Whereas the pipelines related to fishing vessels and passenger ships from Case Study 2 indicated better performance with SMOTETomek transformation, Cargo ships and Case Study 1 did not improve according to their milder degrees of uneven distributions. Consequently, the transformation step was not applied within these models, and the original records were preserved.

Both *RandomForestClassifier* and *XGBClassifier* are provided with long lists of default configured hyperparameters. These are not guaranteed optimal for the given problem and should be tuned explicitly. However, the absolute best hyperparameters are impossible to determine in a decent time. The process requires a trial-error-based approach that can emerge extensively with all combinations within large sets. Therefore, it is common to only tune the hyperparameters considered as having the greatest impact on the learning process. According to Pedregosa et al. (2011) and Chen and Guestrin (2016) the most important hyperparameters and their respective impact on the classifiers are presented in Table 4.7.

Hyperparameter tuning was performed using the Scikit-library functions *Kfold* for cross-validation and the search spaces *RandomizedSearchCV* and *GridSearchCV*. The former method randomly selects samples from a wide grid of predefined hyperparameter values

Hyperparameter	Description
RandomForestClassifier	
n_estimators: {100, 200, 300, 400, 500, 600}	Number of trees in the forest. The default is 100 More trees improve learning but increase time complexity
max_depth: {5, 12, 19, 26, 33, 40}	Maximum number of levels in each decision tree. The default is "None", i.e., splitting will continue until all data belong to one class. Deeper trees are prone to overfit
min_samples_split: {2, 5, 8}	Min. number of samples to be considered when splitting an internal node. The default is 2. Small values may lead to overfitting
min_samples_leaf: {1, 2, 4}	Minimum number of samples allowed in a leaf node. The default is 1. Small values may lead to overfitting
max_features: {2, 3, "auto"}	Maximum number of features considered when performing a split. The default is equal to the total number of features in the data ("auto"). High values may lead to overfitting
XGBClassifier	
n_estimators: {100, 200, 300, 400, 500, 600}	Number of trees in the tree growth. The default is 100. More trees improve learning but increase time complexity
max_depth: {5, 8, 11, 14, 17, 20}	Maximum number of levels in each decision tree. The default is 6. Deeper trees are prone to overfit
min_child_weight: {1, 2, 5, 8}	Minimum sample size threshold for further partitioning. The default is 1. Small values may lead to overfitting
colsample_bytree: {0.4, 0.7, 1}	The fraction of features to be considered when constructing each tree. The default is 1. Exclusion of features when generating trees can prevent overfitting.
learning_rate: {0.05, 0.1, 0.15, 0.25}	Weighting factor used in the update of the next boosting step. The default is 0.3. High values may lead to overfitting
gamma {0.0, 0.15, 0.3}	Weighting factor of minimum reduction in the loss function required to make a further split in the tree. The default is 0. The larger the value, the more conservative the model

Table 4.7: The selection of hyperparameters exposed for tuning and their corresponding initial value ranges

and performs cross-validation of each combination. The combination that provides the best model performance from cross-validation is returned as the most optimally tuned choice. Randomized search does not attempt to try all combinations within the predefined search space and was therefore considered a suitable approach for identifying initial estimates of optimal hyperparameters. As presented in the curly brackets in Table 4.7, candidate values of wide ranges for Randomized search were defined for each hyperparameter. The defaults were used as a starting point for selecting the values within each range, complemented by recommendations from the libraries' documentation and experiences from the ML community platform, Kaggle (Banarjee 2020; Mohit 2020). Next, the outputs from Randomized search were used to narrow the search space before Grid search was applied to the narrowed value ranges for further tuning to the final, optimal configurations. Grid search evaluates *all* hyperparameter combinations instead of sampling in a random manner and is therefore computationally more extensive (Kuhn and Johnson 2019). Three folds of cross-validation were applied in both search spaces to validate the different combinations. The low number of folds was chosen according to the execution time required to run the algorithm caused by the large amounts of data in the search spaces.

The individual outputs from Grid search for each classifier defined the final hyperparameter values applied. The models were individually trained on the training set with their respective hyperparameters using the *fit* method from the Scikit-learn API (Pedregosa et al. 2011). The *XGBoostClassifier* supports an early stopping technique that stops the *fit* function from training when the loss reduction does not improve for a certain number of iterations. This number is provided as input for the classifier and may overwrite the value of the *n_estimators* hyperparameter if the model performance does not improve. Hence, early stopping ensures that the model does not overfit due to an unnecessary amount of constructed trees. The stopping criterion was defined by observing the evolution in loss reduction during model learning. In all cases, the loss did not reduce further after being stable for ten iterations, and the stopping criterion was set as such.

4.3.3 Model evaluation

Although the resampling approach increased the presence of the minority classes, the training data were not fully balanced. As explained in Section 2.8.2, evaluation by the standard accuracy metric on unevenly distributed data may mislead the actual model performance. Therefore, several performance metrics were applied to assess the models correctly. As confusion matrices support evaluating each class irrespective of class proportions, the precision,

recall, and F1-score were used, in addition to accuracy, as the primary performance metrics for cross-validation during model training and the final evaluation on the test set.

4.3.4 Overview of model development

Figure 4.13 summarizes the steps described in this chapter for applying the RF and XGBoost classifiers to the pre-processed data presented in Section 4.2.8.

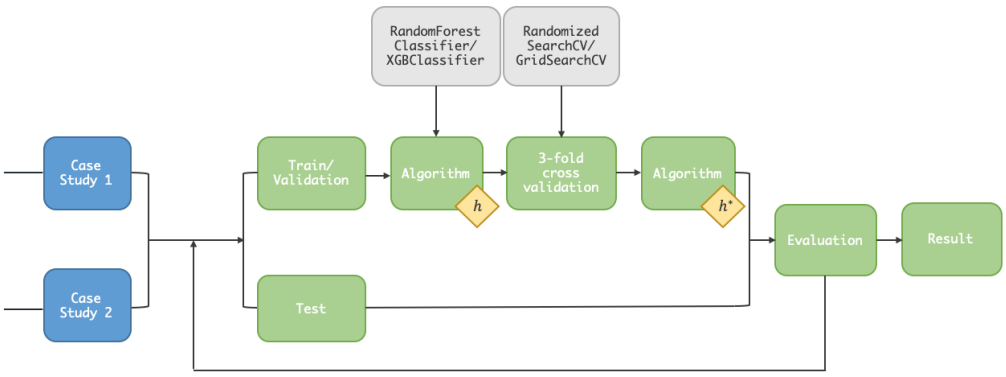


Figure 4.13: Flowchart of the ML development phase. h and h^* refer to default and optimized hyperparamters, respectively

Chapter 5

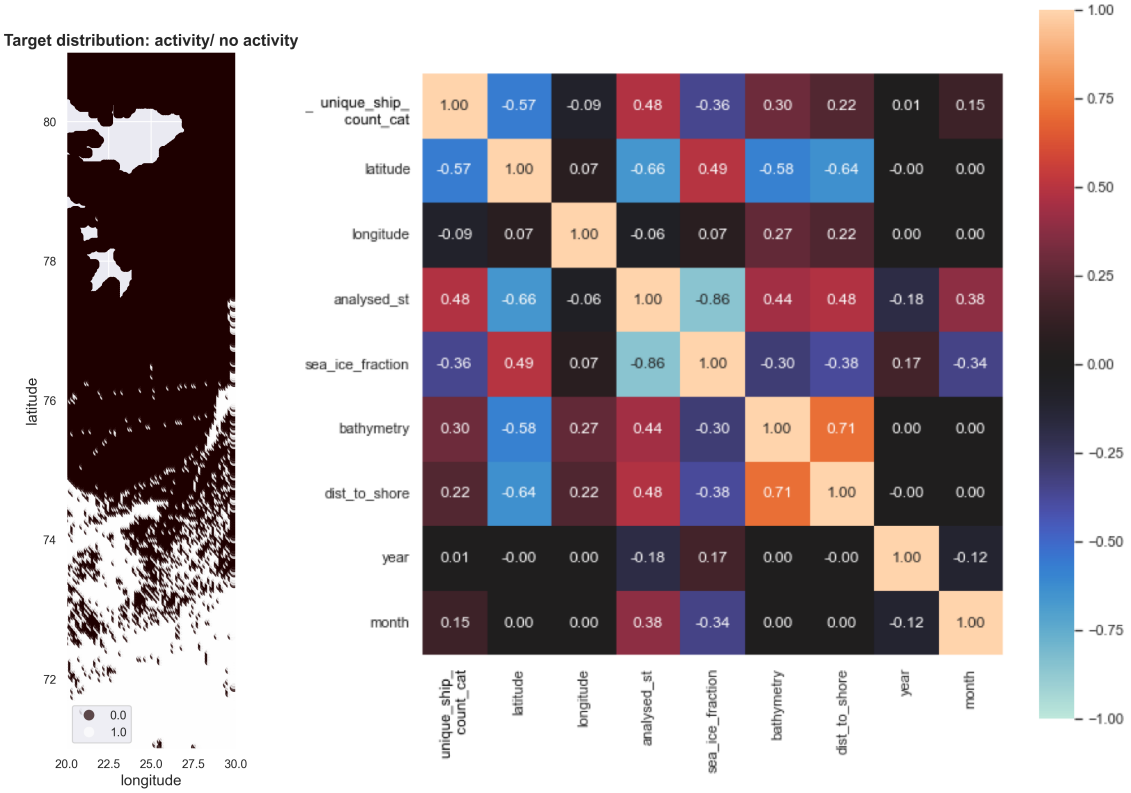
Results

The presented case studies aim to demonstrate and validate how to apply supervised ML for identifying activity patterns in an Arctic area based on surrounding explanatory variables describing the environment and ecology. This section presents the results from the case studies and provides the foundation for validating the proposed methodology. The ensemble classifiers' tuning, learning performance, and predictive capabilities are presented in light of each target definition.

5.1 Case study 1: Predicting presence or absence of activity

Case Study 1 targets vessel presence and absence in the period from 2015 to May 2021 within an area of the Barents Sea, spanning from the Norwegian coast above latitude 71° to 81° and longitude 20° to 30°. Figure 5.1a presents the spatial contrasts between absence (0) and presence (1) of activity within all years of the period, viewed as black and white points, respectively. The upper left areas in light grey correspond to the Svalbard islands. The plot indicates that activity presence stands out in the southern latitudes. Although there are cases of presence further north, these are dominated by the absence points within the whole time frame; hence, these cases are not directly visible.

Figure 5.1b shows a heat-map describing the relationship between the features of Case Study 1, where the target is represented by *unique_ship_count_cat*. Each pair of features is connected by their Pearson product-moment coefficient, representing their linear correlation by a value between -1 and 1. The higher the absolute covariance value, the more correlated the features are. The heat map shows that *latitude* is the target's most correlated feature by a value of -0.57, meaning that higher latitudes correspond to less activity. This confirms the southern



(a) Spatial plot of target distribution

(b) Heat map of feature correlations

Figure 5.1: Case Study 1 - Target distribution

dominance of activity presence viewed in Figure 5.1a. *analysed_st* (SST) is the next most correlated (0.48), followed by *sea_ice_fraction* (SIC) (-0.36) and *bathymetry* (0.30). Their respective values indicate that vessel presence is more likely with higher temperatures and, to some extent, less sea ice and deeper waters. Whereas *dist_to_shore* (distance) and seasonal variations by *month* show less correlation, yearly changes and activity presence have no significant relationship.

The internal correlations between the predictors are also presented in Figure 5.1b, of which the relationships between the SST and SIC features (-0.86) and the bathymetry and distance features (0.71) stand out the most. The former yields that higher temperatures reduce the ice concentration, whereas the latter yields that the distance to the nearest coast increases with the underwater depth. According to the earth's natural physics, these deductions are reasonable, which justifies the predictors' credibility. Moreover, both SST and SIC have higher seasonal correlations by month than year. Their coefficients by year of -0.18 and 0.17 are

rather small. However, they indicate that the sea surface temperature decreases and the sea ice increases by year.

5.1.1 Hyperparameter optimization

Table 5.1 illustrates the outcome of sequentially applying *RandomizedSearchCV* and *GridSearchCV* to each classifier for hyperparameter optimization. The classifiers were validated before and after applying the generated hyperparameters from each search space. Hence, the percentages yield the corresponding improvement in validation score from replacing the default hyperparameters with the tuned values. The models were validated by balanced accuracy, which yields the average evaluation of recall obtained on each class (Pedregosa et al. 2011).

Hyperparameter tuning by the search spaces improved the validation scores of both models, especially XGBoost, where the final optimization increased the validated accuracy by 2.32%. However, the effect of further narrowing the search space and applying *GridSearchCV* did not improve the RF performance and only slightly increased the validated accuracy of XGBoost.

	RandomForestClassifier	**XGBClassifier**
RandomSearchCV	0.30%	2.12%
GridSearchCV	0.27%	2.32%
Final hyperparameters	n_estimators: 200 max_depth: 40 min_samples_split: 2 min_samples_leaf: 2 max_features: 3	n_estimators: 250 max_depth: 25 min_child_weight: 5 colsample_bytree: 1 learning_rate: 0.05 gamma: 0.1

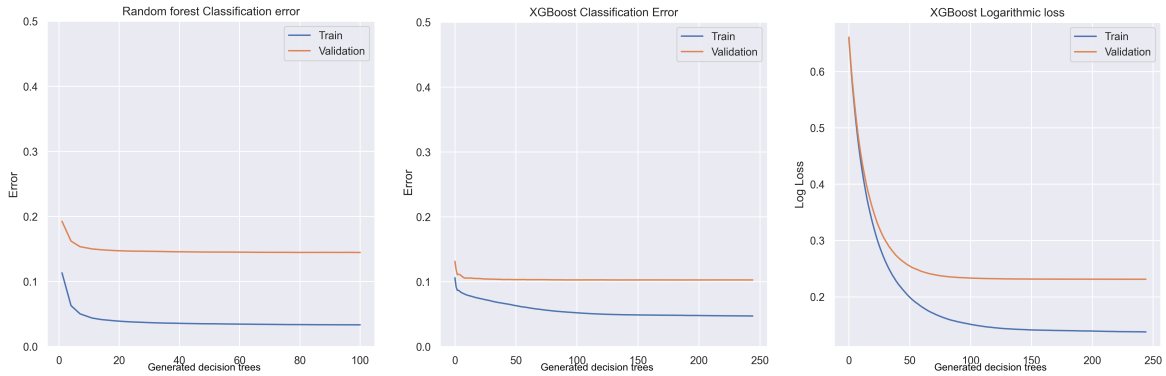
Table 5.1: Case Study 1 - Hyperparameter tuning: improvement in balanced accuracy by replacing the default hyperparameter values with tuned values from each search space

The tuning process proved that Randomized search tends to choose high values of *n_estimators* and *max_depth*. For example, it generated 500 trees (estimators) and "*None*" depth for the RF classifier. A high number of trees improves performance but may slow down the training process. Similarly, deeper trees capture more information but are prone to overfitting due to increased specificity on particular samples (Mitchell 1997). Hence, to increase model simplicity, the value ranges for *n_estimators* and *max_depth* were manually reduced before applying Grid search, despite the high outputs from Randomized search. The finally optimized XGBoost classifier resulted to be more risk-averse than RF according to the small maximum depth, the small learning rate, and the high gamma value. Additionally, it required a high sample

threshold for further partitioning, while RF had a low number of minimum leaf and split samples.

5.1.2 Learning curves

Insight into the classifiers' learning performance was obtained by continuously cross-validating the training process. The learning curves in Figure 5.2 express this performance by plots of the evaluation metrics on the y-axis along with the number of generated trees on the x-axis. The RF and XGBoost classifiers in Figure 5.2a and 5.2b, respectively, are evaluated by classification error, which yields the percentage of misclassified labels in the data. The remaining plot in Figure 5.2c shows XGBoost evaluated by logarithmic loss.



(a) RFClassifier: (balanced) classification error

(b) XGBClassifier: classification error

(c) XGBClassifier: logarithmic loss

Figure 5.2: Case Study 1 - Learning curves by number of generated trees

The blue and orange curves in Figure 5.2 correspond to the train and validation learning curves, respectively. The former yields model evaluation on the training data and expresses how well the model "learns" the given observations. The latter corresponds to model evaluation on a hold-out validation set from three-fold cross-validation and describes how well the model "generalizes" to new data. As seen, the evaluation metrics score better on the training data, as these observations are familiar to the model. Both curves within each plot evolve similarly and decrease with time towards relatively small error values meaning that the models improve with experience. However, there is a notable gap between the curves. Enhanced learning stops when the curves reach their respective plateaus. Whereas the error curves converge quite fast at around 20 decision trees, the logarithmic loss decreases toward 100 decision trees.

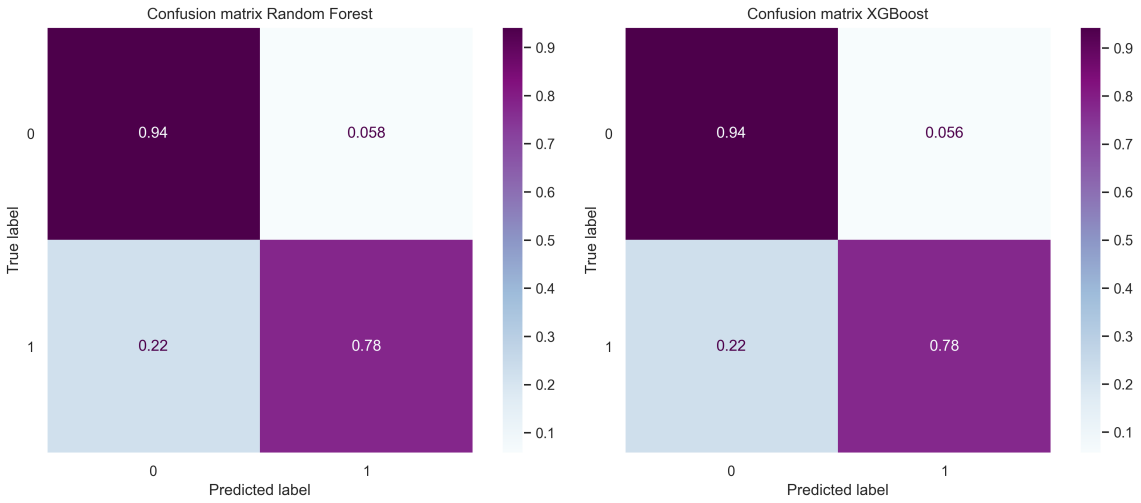
5.1.3 Prediction performance

20% of the initial data, equivalent to 330885 records, was used as a test set for a final evaluation of the models' prediction performances. Both classifiers performed identically by the evaluation metrics provided for each class in Table 5.3. The class of most support, i.e., *absence*, obtained high values of both precision and recall, summarized by an F1-score of 93%. This means that the models managed to predict almost all absence labels correctly as well as captured most of the absence labels that actually are absence. On the other side, prediction of *presence* obtained a slightly better precision score than recall, and an F1-score of 80%, which indicates that the models were better at correctly predicting presence than identifying *all* actual presence cases in the data.

	Precision	**Recall**	**F1-score**	**Support**
Class 0 (absence)	0.91	0.94	0.93	238252
Class 1 (presence)	0.84	0.77	0.80	92633
average	0.88	0.86	0.87	
accuracy			0.90	

Table 5.3: Case Study 1 - Classification report of prediction metrics (similar results from both classifiers)

The model performance described above can be summarized for each classifier by the confusion matrices presented in Figure 5.3. The values are normalized by the number of true labels within each class, meaning that the values of the diagonal correspond to the recall values from Table 5.3. Although the overall accuracy proves that the models correctly predicted 90% of all samples, the confusion matrices reveal that 22% of the actual presence values were incorrectly classified as absence. However, the overall performance given by the F1-score average of 87% does not differ significantly from the accuracy.



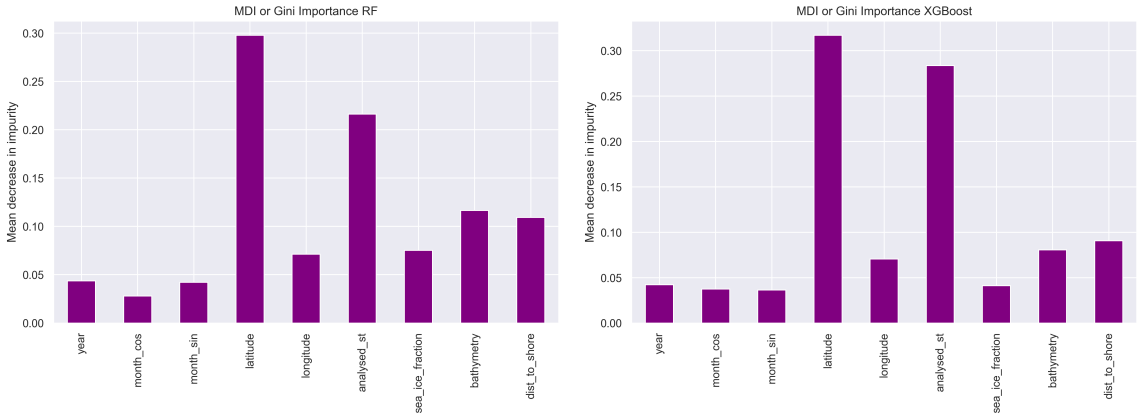
(a) RFClassifier (b) XGBClassifier

Figure 5.3: Case Study 1 - Prediction performance

5.1.4 Feature importance

Feature importance refers to which features the ensemble models prefer when generating the individual decision trees. These features are selected based on their expected contribution toward model interpretation against new data, given by their information gain or impurity decrease, as described in Section 2.6.2. The importance is calculated for each decision tree, and the total value across all trees of the ensemble model yields the final feature importance (Brownlee 2020).

Figure 5.4 shows the normalized importance values for the RF and XGBoost classifiers. The plot shows that *latitude* is the most important feature, followed by *analysed_st* (SST), *bathymetry*, and *dist_to_coast* (distance). On the other side, *longitude*, *sea_ice_fraction* (SIC), and the time-encoded features are those of less model preference. The order of important features corresponds well to the target's coefficients in absolute values presented in Figure 5.1. However, the feature importances of RF have a more even distribution than XGBoost, where latitude and SST stand out significantly compared to the rest.



(a) RFClassifier (b) XGBClassifier

Figure 5.4: Case Study 1 - Feature importance

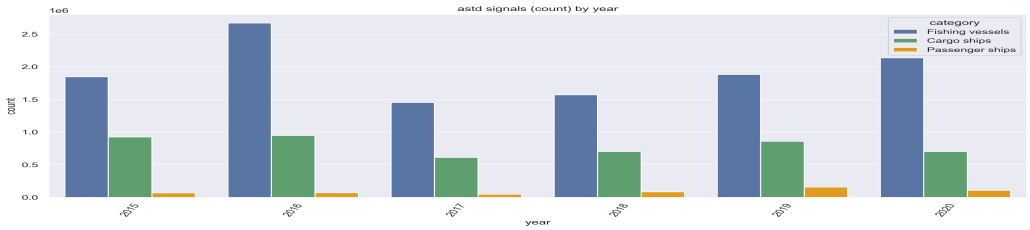
5.2 Case study 2: Predicting activity density by industry

This section presents the results from the second case study, which aims to estimate vessel density by three classification levels within one of the industries: fishing, cargo shipping, and tourism. The case study considers cases of vessel presence, meaning that the region bounded by latitudes 71° and 81° and longitudes 30° to 20° is further restricted to the spatial grids where at least one ship is observed. The time frame ranges from 2015 to May 2021, except for the fishing industry, which is upper bounded by year-end 2019. Whereas fishing vessels constitute the highest share of AIS messages, the total number of individual Cargo ships is higher.

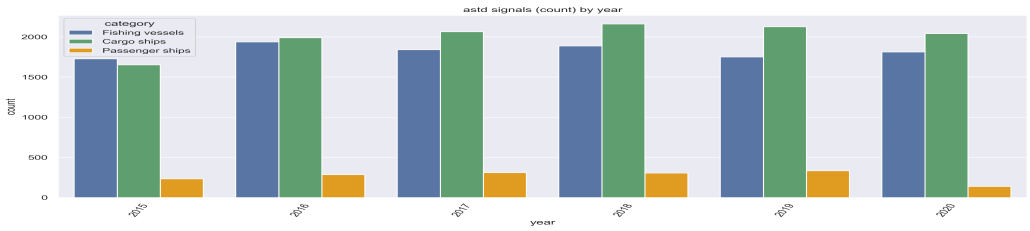
5.2.1 Target distribution

Figure 5.5a shows the evolution in the number of all AIS transmissions within the period, except from 2021, as only half of the year is represented. Figure 5.5b shows a similar distribution excluding multiple messages from the same vessel. It is observed that the number of AIS messages drops from 2016 to 2017 before it tends to increase toward 2020. However, the number of unique ships transmitting the messages does not change significantly from year to year.

Figure 5.6 shows each industry's monthly numbers of AIS transmissions. The y-axis' upper limit on the individual plots changes by industry; hence, fishing vessels stand out in the total number of transmissions. Whereas Figure 5.6a and 5.6b match the yearly evolution expressed in Figure 5.5, the distribution of messages from passenger ships in Figure 5.6c shows

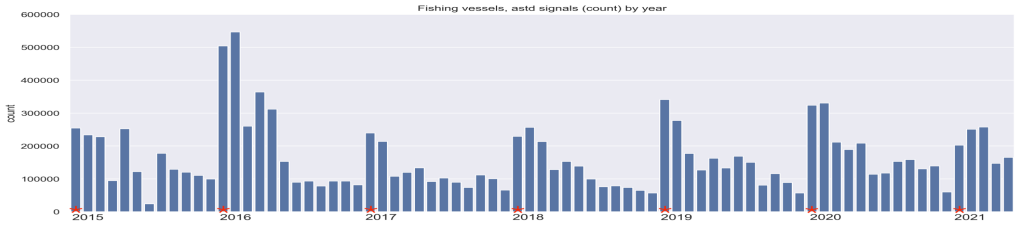


(a) AIS messages by industry

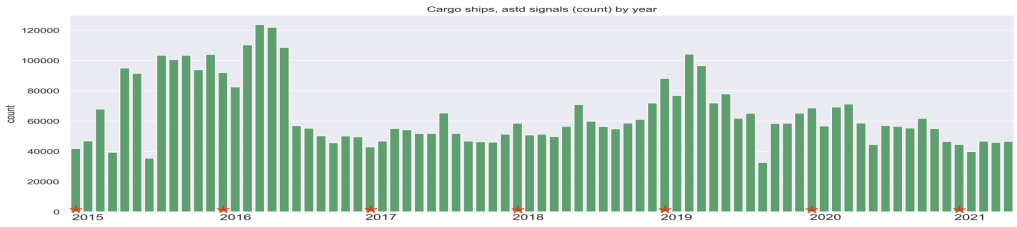


(b) Number of unique vessels by month

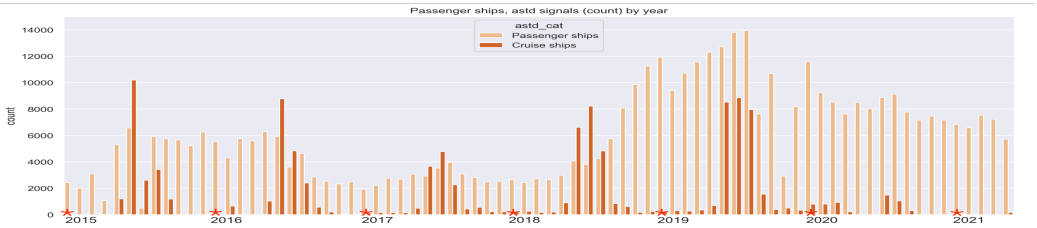
Figure 5.5: Case Study 2 - Yearly distributions of ASTD data by industry



(a) Fishing vessels



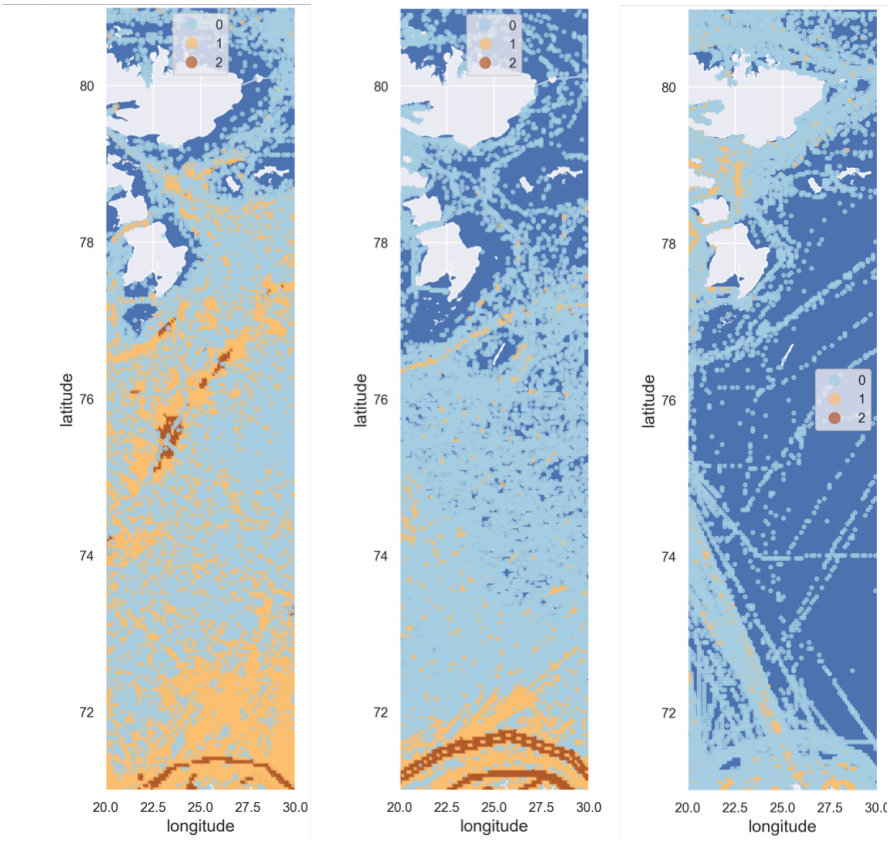
(b) Cargo ships



(c) Passenger ships

Figure 5.6: Case Study 2 - Monthly distributions of AIS messages by industry. The red marks indicate the beginning of each year (January). Note that the ranges of the y-axis differ according to industry

a significant increase from year-end 2018. The plot is distinguished by regular passenger ships and cruise ships, as defined by the ASTD Level 3 category, and reveals a high seasonal trend in cruise ship activity. As seen, the summer season of 2020 contains far fewer cruise ship messages than the rest. In contrast, the seasonal variations in AIS messages among the smaller passenger ships are less clear.



(a) Fishing vessels (b) Cargo ships (c) Passenger ships

Figure 5.7: Case Study 2 - Spatial distributions of target values. The dark blue background area corresponds to the sea

Figure 5.7 presents the spatial target class distributions within each industry. Cargo ships constitute the most obvious distinction in ship density level by space, mainly concentrated at the lower latitudes where the pattern of the shipping route stands out significantly by the highest density class (Class 2). Fishing vessels, however, have an arbitrary distribution, spread over the region. Lastly, the densities of passenger ships resemble straight-line patterns across the deep sea but tend to stay near the coast. A combined overview of the industries' mean activity distribution by time and space is provided in Figure 5.8. The plot indicates

that cargo vessels remain at the lower latitudes, while fishing and tourism activities further north have a seasonal variation. Moreover, there is a slight increase in tourism activities during summer at higher latitudes from 2015 to 2018. However, the line plot does not reveal any overall changes in seasonal variations over time.

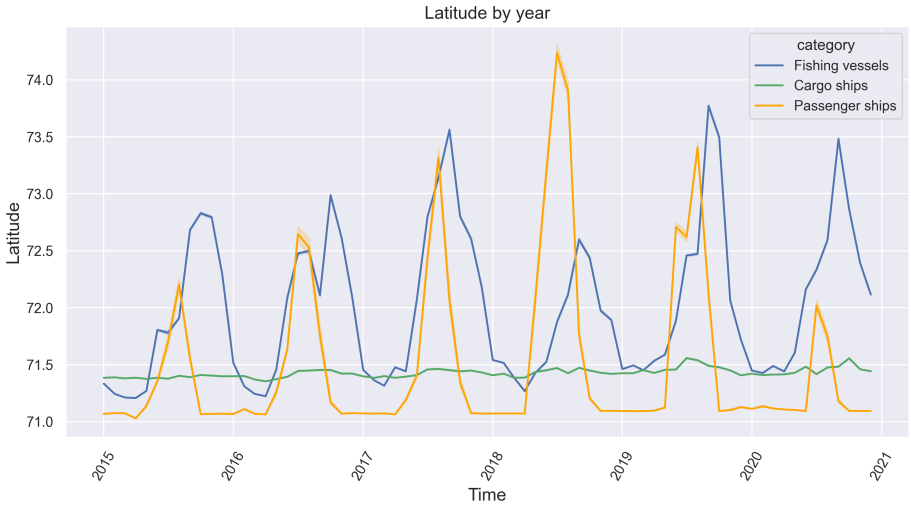


Figure 5.8: Case Study 2 - Evolution in mean positional latitudes from AIS records by industry

5.2.2 Hyperparameter tuning

Similar to Case Study 1, hyperparameter tuning proved more valuable for the XGBoost model than RF. Table 5.5 presents an overview of the results from the tuning process, where the improvement in balanced accuracy corresponds to the final output from *GridSearchCV*. The hyperparameter tuning of RF did not improve the accuracy, of which changes are essentially negligible. However, the XGBoost classifier improved within all industries. The table shows that the fishing industry stands out the most, with a 5.96% improvement compared to the default model. In contrast, the accuracy improvement of the passenger ships classifier is minor.

It is observed that the classifiers of fishing vessels and cargo ships have the most similar outputs of finally tuned hyperparameters. The XGBoost classifier of passenger ships preferred a relatively low number of trees and a small depth, by 50 trees and 15 levels, respectively, and both classifiers applied all possible features when splitting the trees. In contrast, the hyperparameters of the fishing and cargo classifiers involved smaller samples of features in each split, as well as lower learning rates and higher sample size thresholds for further partitioning with XGBoost.

	RandomForestClassifier				**XGBClassifier**		
	Fishing	Cargo	Tourism		Fishing	Cargo	Tourism
Percentage improvement	-0.15%	0.13%	-0.11%		5.96%	3%	0.82%
Hyperparameters							
n_estimator:	400	400	350	n_estimators:	200	400	50
max_depth:	30	40	25	max_depth:	20	25	15
min_samples_split:	4	3	6	min_child_weight:	8	5	3
max_features:	2	3	"auto"	colsample_bytree:	0.8	0.8	1
min_samples_leaf:	2	2	1	learning_rate:	0.05	0.05	0.15
				gamma:	0.05	0.0	0.0

Table 5.5: Case Study 2 - Outcome of hyperparameter tuning

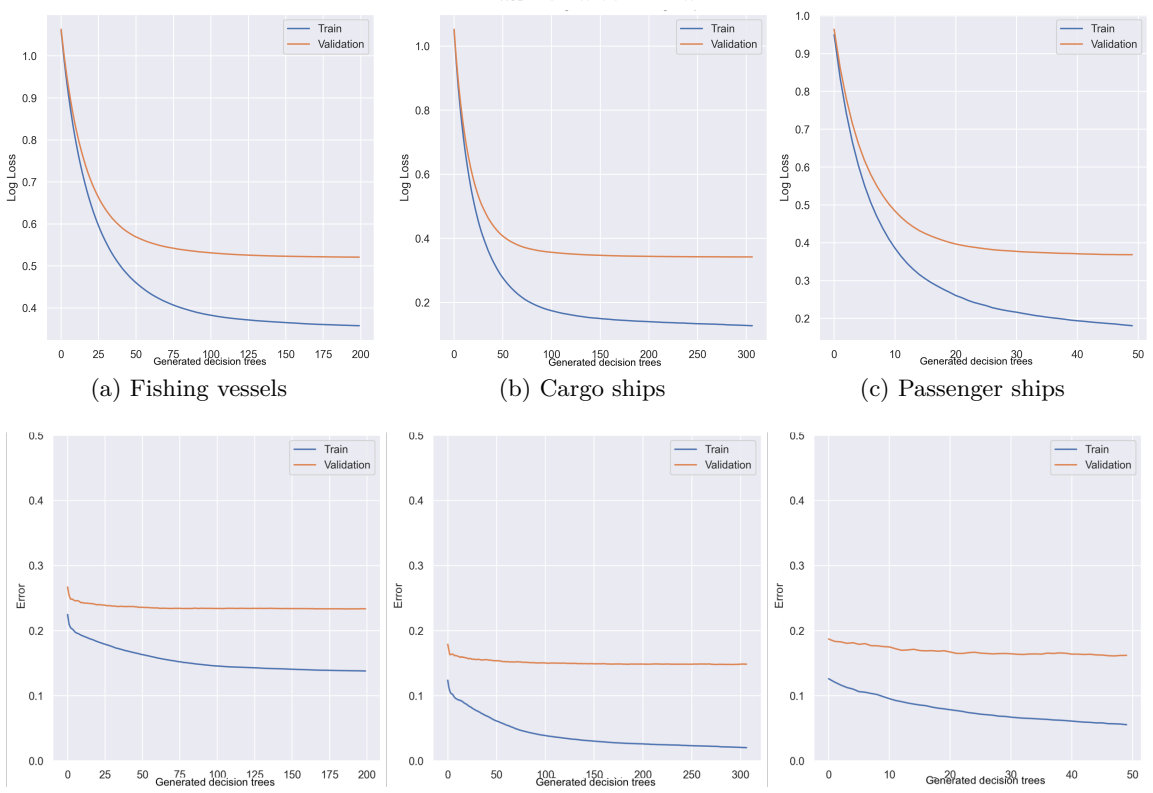
5.2.3 Learning curves

The ensemble classifiers' learning processes within each data set by industry evolved rather similarly. Subsequently, they performed relatively equivalent, such as in Case Study 1, except in some of the less supported class predictions where XGBoost outperformed RF by minor percentages in precision and recall. Therefore, the remaining part of this chapter will focus on the learning process and prediction performance of the *XGBoostClassifier*.

The learning curves of Case Study 2 are presented in Figure 5.9, where the y-axis corresponds to the score of the evaluation metrics along with the number of generated trees on the x-axis. The plots of the logarithmic loss function by industry in the upper level of the figure prove that the classifier increases certainty by the number of decision trees generated. Whereas the validation loss from fishing vessels stabilizes at 0.52, the classifier performs slightly better against the cargo ships and the passenger ships by values of 0.34 and 0.37, respectively. Correspondingly, the plots in the lower level of Figure 5.9 prove that cargo ships and passenger ships outperform fishing vessels by lower classification errors. Furthermore, the curves representing passenger ships converge faster than the others, at around 50 boosting rounds, which corresponds well to the output from the hyperparameter tuning. Although the curves behave stably, similar gaps as observed in Case Study 1 also apply to the industries in Case Study 2.

5.2.4 Prediction performance

The train-test split resulted in test sets of 61398, 28022, and 5777 records of fishing vessels, cargo ships, and passenger ships, respectively. Table 5.6 presents each XGBoost classifier's performance accuracy, class evaluation metrics, and their individual support. It should be



(a) Fishing vessels (b) Cargo ships (c) Passenger ships

(d) Fishing vessels (e) Cargo ships (f) Passenger ships

Figure 5.9: Case Study 2 - Learning curves of XGBoost by number of generated trees

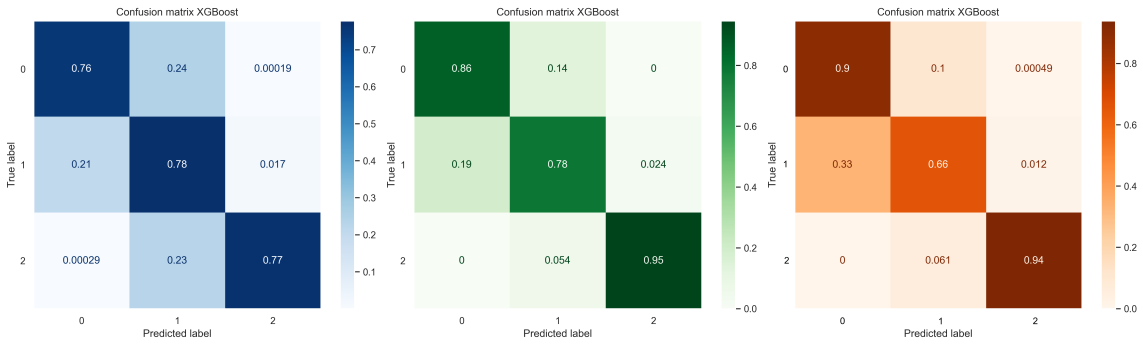
noted that the support values correspond to the initial class proportions before resampling my SMOTETomek, as the test set preserved data originality to obtain a proper evaluation of the model performance. The values of the table indicate that the classifiers performed promising and somewhat similarly against cargo and passenger ships with an overall accuracy of around 85%. However, the prediction of fishing vessel class densities obtained the lowest accuracy and F1-score of 77%.

The recall value in this context corresponds to the proportion of instances within each density level that the classifier managed to detect, while the precision yields the proportion of correctly labeled density classes from all predictions made. All classes within the three industries obtained close precision and recall values which means that the classifier predicted accurately and detected relevant cases to a similar extent. Consequently, the weighted average of the two metrics, i.e., the F1-score, did not differ either. An interesting observation is that the F1-score yields the best result by the minority class for all industries, of which cargo ships and passenger ships obtained high scores of 95% and 93%, respectively.

		Accuracy	Precision	Recall	F1-score	Support
Fishing vessels	Class 0	0.77	0.76	0.76	0.76	26754
	Class 1		0.77	0.78	0.77	31142
	Class 2		0.84	0.77	0.80	3502
Cargo ships	Class 0	0.85	0.85	0.86	0.86	12388
	Class 1		0.80	0.78	0.79	10103
	Class 2		0.96	0.95	0.95	5531
Passenger ships	Class 0	0.84	0.88	0.90	0.89	4057
	Class 1		0.68	0.66	0.67	1442
	Class 2		0.93	0.94	0.93	278

Table 5.6: Case Study 2 - Classification report of prediction metrics by industry

The confusion matrices in Figure 5.10 provide further visual insight into which density levels the misclassifications are related to. The matrices indicates that distinguishing between Class 0 (low) and Class 1 (medium) densities was the most challenging for the classifiers. This is viewed by the high fractions of Class 0 predictions which actually belong to Class 1, and the corresponding fractions of Class 1 predictions which actually are Class 0. For example, 33% of the incorrect predictions of Class 0 in Figure 5.10c actually belong to Class 1, meaning that the classifier struggled with distinguishing these two classes. Conversely, Class 2 (high) was less misinterpreted than the other classes, except for the fishing vessel classifier, where 23% of the Class 1 predictions were actually Class 2.



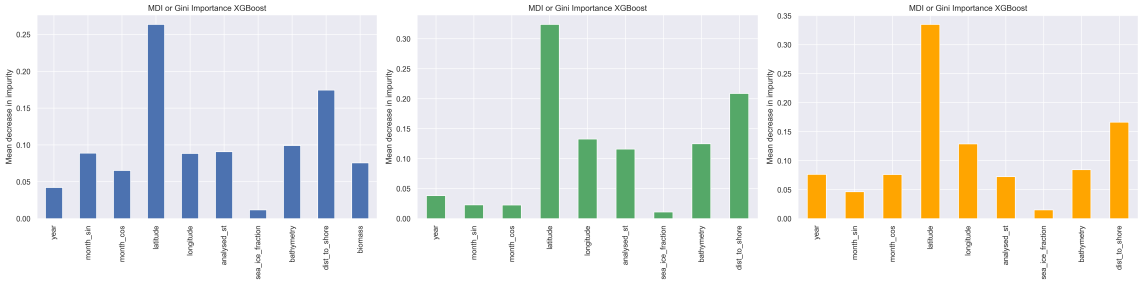
(a) Fishing vessels (b) Cargo ships (c) Passenger ships

Figure 5.10: Case Study 2 - Normalized confusion matrices by industry

5.2.5 Feature importance

It is evident from Figure 5.11 that the outstanding feature of importance continues to be *latitude*. Additionally, *dist_to_coast* (distance) to the nearest coast is the second most important feature within all industries. Hence, the classifiers primarily depended on the spatial

features for prediction. *analysed_st* (SST) is approximately equally moderately important for all industries. On the other hand, the *sea_ice_fraction* (SIC) attribute remains the least popular attribute.



(a) Fishing vessels (b) Cargo ships (c) Passenger ships

Figure 5.11: Case Study 2 - Most important features for each classifier given industry

The classification tasks of the fishing vessel and passenger ship densities depended more on the temporal features than the classification of cargo ship density, which is further justified by the target distributions provided in Section 5.2.1. Furthermore, 5.11c indicates that the classifiers identified *bathymetry* as less important for passenger ships compared to the others. Figure 5.11a shows no clear linkage between *biomass* and the prediction of fishing vessel densities. This is further described by the correlation plot in Figure 5.12, which also shows that the number of unique vessels by year has decreased relative to biomass.

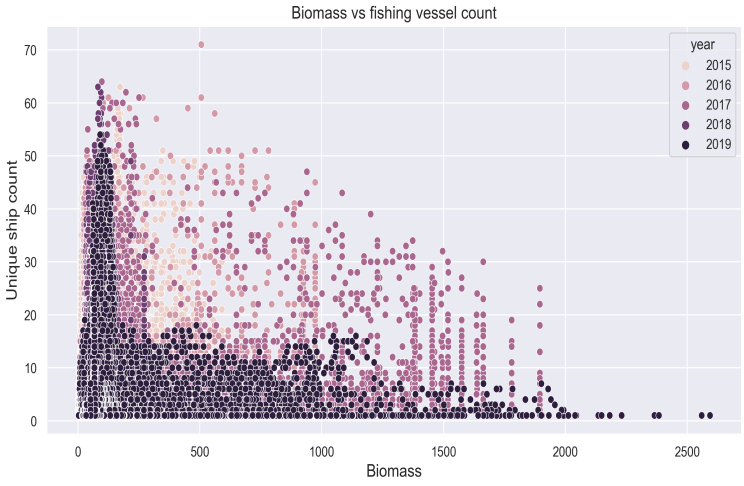


Figure 5.12: Case Study 2 - Correlation plot between biomass and unique ship count of fishing vessels (i.e., target before binning)

Chapter 6

Discussion

In this chapter, the ML learning processes and outcomes of the case studies are compared and evaluated, and the predictors' validity is assessed. Finally, a discussion of the proposed solution's applicability is provided.

6.1 Model performance review

According to experiences presented in Section 2.7 and 4.3.1, it is expected that XGBoost would outperform RF by its ability to minimize the error from the predecessor tree. The benefit from this optimization step is primarily observed by Case Study 2, where XGBoost behaved more offensively than RF in capturing the minority classes of activity density prediction by industry. Compared to XGBoost, the RF classifier could not identify the classes of low participation similarly since it builds the decision trees independently of each other. However, except for the few additional correctly labeled classes by XGBoost, the results show that the classifiers performed somewhat similarly during learning and that the overall final prediction outputs did not diverge significantly. The common behavioral patterns can be related to the fact that the classifiers share the same foundational building block, the decision tree, and target the features similarly.

6.1.1 Model learning

As expressed in Section 4.3.1, a higher impact of tuning XGBoost is expected according to the classifier's sensitivity to hyperparameters. This is further confirmed by Table 5.1 and 5.5, which prove that XGBoost better benefited from hyperparameter tuning than RF in both case studies. Furthermore, the minor changes in the RF performances can be explained by the

narrowing of value ranges after Randomized search, which potentially led to too conservative combinations of hyperparameters values compared to RFs preference for high *n_estimators* and *max_depth* values. However, such a narrowing was considered necessary to reduce time complexity according to the large amounts of data and prevent overfitting caused by perfectly learned training data from unnecessary splits in deep trees. Moreover, overfitting tendencies are not observed from the learning curves as both the train and validation curves have similar behavior without sudden discrepancies. The early stopping technique in XGBoost ensured that the training process stopped before the validation error trended from descending to ascending.

A comparison of the results from hyperparameter tuning by industry in Table 5.5 proves that the classifiers against the fishing and cargo industries had more similar outputs from tuning compared to passenger ships. The reason for this can be related to the respective large data sets of fishing vessels and cargo ships as well as their class distributions, which spatially tend to evolve similarly, as seen in Figure 5.7. On the other hand, the passenger ships data contain fewer records, which led to fewer trees and less depth. In addition, its optimized XGBoost hyperparameters differed less from the default values stated in Table 4.7, which may explain the minor percentage improvement after hyperparameter tuning compared to the other industries.

As observed in Figure 5.2 and 5.9, the validation and train curves behave similarly and converge for both case studies. However, the noticeable gap between the curves may indicate that the training data did not provide adequate information relative to the validation data set. This can be explained by the uneven target distribution not being considered when extracting the hold-out validation set. Alternatively, the unrepresentative data is a consequence of only using three folds in cross-validation, causing the proportion of the validation set to be too large relative to the remaining training set. The fact that the learning process was affected by the uneven class distributions can be justified by the fast convergence in the classification error curve relative to the logarithmic loss, which is visualized by fewer tree generations on the x-axis. Whereas the former expresses the percentage of incorrectly predicted labels, the latter yields the likelihood of correct predictions, i.e., how confidently the XGBoost model predicts the class of a given record. Hence, the fast convergence in classification error might be affected by uncertain guesses that provided the correct output according to the majority class, as elaborated in Section 2.8.2. However, according to the logarithmic loss, it is evident that more training is required to improve model certainty.

6.1.2 Predictive capabilities

As presented in Table 5.3, the final evaluation from the test set of Case Study 1 yields better model performance in predicting the absence class, which is the class of most support in the data. Although the data set has a mild degree of imbalance by a 72%/28% ratio of the absence and presence classes, respectively, the high F1-score of the majority class relative to the minority indicates that the final prediction performance was affected by confusion from the uneven class distribution. However, despite the imbalance, the model performed satisfactorily, proven by the averaged F1-score of 87% being only minor percentages below the 90% accuracy.

In contrast to Case Study 1, the classifiers provided the best predictive performance on the minority classes of Case Study 2. This performance can be a result of the SMOTETomek resampling, as explained in Section 4.2.7, which was not applied on the test set to maintain the original distribution, and evaluate the performance properly. Hence, the support values in Table 5.6 correspond to the initial class proportions. As previously justified in Section 4.2.7, resampling was not performed on the data of cargo ships in Case Study 2 nor the data of Case Study 1, which further explains the high attraction to the absence class in the latter case. However, although the minority class support of cargo ships is higher compared to the other original data sets by industry, the corresponding models' satisfactory results can be related to the clear spatial separation between the cargo ship density classes, as seen in Figure 5.7b. The overall distributions from Figure 5.7 may further explain why the classifiers of Case Study 2 struggled with distinguishing between Class 0 (low) and Class 1 (medium), as these have less conspicuous patterns compared to Class 2 (high).

Overall, the relatively high F1-scores and the small fluctuations in precision and recall within each classification task of the case studies prove that the classifiers managed to detect patterns and predict activity-related targets by external predictors assembled from different sources. However, although RF and XGBoost were considered suitable learners according to this thesis's purpose and scope, there exist several potential classifier candidates that most certainly would have obtained similar results as these ensemble learners. As the field of ML evolves continuously, there will constantly be developed more advanced methods that outperform others in either evaluation metrics, time complexity, or even both. However, instead of comparing different classifiers by their prediction performances, this thesis aims to illustrate the feasibility of applying such learners in an activity-predictive context focused on the Arctic region. Hence, the fact that both the RF and XGBoost classifiers provided close to similar and relatively high evaluation scores, despite imbalance, justifies the performance of

one another, not to mention the possibilities such supervised learners have to offer in this context.

6.2 Assessment of the predictors

Figure 5.4 shows that the classifiers' respective proportions of feature importance differ, which can be explained by how the classifiers build their trees. Two well-correlated features may contain similar information, hence have equal classification importance (Chen and Guestrin 2016). Since XGBoost generates the trees in sequence, it will choose one feature among several correlated and tend to apply this specific learned link between the selected feature and the outcome along the learning process. On the other hand, RF builds each tree independently from the others and may choose randomly among several correlated features for each tree generation. This is why the features were frequently applied in RF and provided more overall importance.

Although the RF importance values were more evenly distributed than XGBoost, the two classifiers had similar orders in feature preference, which is reasonable as both learners construct decision trees by information gain. The overall performance presented by the averaged F1-score proves that the classifiers performed best in Case Study 1, followed by density prediction of cargo ships, passenger ships, and fishing vessels in Case Study 2. The reason why Case Study 1 was the most predictable can be explained by the fact that only two classes of different distributions were considered, while the latter involved an extra class which further contributed to imbalance uncertainties. Another explanation regards the fact that the features potentially are more significant in terms of determining *whether* there is activity rather than *how much* activity.

Figure 5.4 shows that the most outstanding features of importance in Case Study 1 are latitude and SST. The order of importances is further justified by the heat map presented in Figure 5.1b, which yields that latitude and SST have the highest correlation coefficients with the target. Whereas the classifiers of Case Study 2 similarly preferred latitude, Figure 5.11 indicates that the climatic factors changing by time were less preferred. In particular, SIC proved to be of the lowest importance, which contradicts the expectation according to experiences from the literature provided in Section 2.4. However, the other spatial features, including bathymetry and distance to the nearest coastline, had higher importance scores. The difference in the SST and SIC preferences can be explained by the fact that Case Study 1 has a higher spatial and temporal coverage than Case Study 2, which led to higher impacts of changes in sea ice extent and temperature according to the Barents Sea Polar Front. In

contrast, Case Study 2 is restricted to the spatial regions where activity presence is a fact, making it reasonable to assume that the associated climatic features are more stable. Another reason why Case Study 2 preferred spatial features over the others can be related to the industries' objectives. For example, the high importance of bathymetry and distance to the coast within the fishing vessels and cargo ships can be connected to the locations and depths of common fisheries and the shipping passage close to the Norwegian coast, respectively. On the other hand, passenger ships had lower bathymetry importance than the others, which is reasonable considering that tourists may prefer staying on shallow seas to observe species and nature. Nevertheless, the outstanding importance of latitude justifies the classifiers' abilities to classify better the density levels that are clearly distinguished by their respective spatial patterns.

An interesting observation is that the ecological aspect included as part of the fishing vessel data did not contribute remarkably to density prediction. Instead, it increased the dimensionality of the fishing vessel data, which may explain why the corresponding evaluation metrics from the prediction of fishing vessel densities are lower than the others. The fact that not all features were useful in this context is further justified by the choice of values of *max_features* and *colsample_bytree* from hyperparameter tuning, provided in Table 5.5. Although the fishing vessel density is expected to increase with higher biomass values, the plot in Figure 5.12 slightly illustrates the opposite. However, as elaborated in Section 4.2.4, the biomass data set was manipulated due to its insufficient resolutions in space. Consequently, the poor ecological significance may be caused by unrepresentative data values generated by interpolation rather than the information the original data aims to convey.

6.3 Activity trends over time

According to the studies presented in Section 2.4, it is expected to see an increasing activity trend over time. Whereas such an increase is not observed by unique ships in Figure 5.5b, Figure 5.5a shows a gradual increase in AIS transmissions before and after the sudden drop between 2016 and 2017. This comparison indicates that the individual vessels operate more frequently by year. For example, while fishing vessels transmit messages more frequently than cargo ships, the number of individual fishing vessels is lower. According to Section 2.3.3, this means that individual fishing vessels operate more often and at higher speeds compared to cargo ships.

The sudden drop in AIS messages can have many reasons. ICES (2018) reported a stock collapse of capelin in 2016 in the Barents Sea. The pelagic fish is an important prey for various

predators, including cod, seals, seabirds, and whale species. Hence, fluctuations in the capelin stock have impacts higher up in the food chain. The capelin collapse in 2016 caused the fishery to temporally close, which is a potential reason for the significant decline in AIS messages from fishing vessels. The fishery restrictions were, however, lifted for 2018, which may justify the subsequent increase. Furthermore, it was reported that 2016 was affected by significantly high temperatures compared to other years and that 2017 was a year of record-high storm activity. These observations may explain the sudden peak in 2016, followed by the subsequent reduction in AIS messages within all industries.

In addition to the distribution in Figure 5.5b, the heat map in Figure 5.1b and the feature importances in Figure 5.4 and 5.11 substantiate a weak relationship between the time-encoded *year* feature and the activity target of both case studies. However, it is worth pointing out that AIS was not compulsory for all vessels until recent years, making it reasonable to question whether the distribution in records is affected by the continuous growth in AIS ground stations, satellites, and vessels having such a system. The unclear temporal relationship may additionally be influenced by third variables such as the Covid-19 breakout, which, for example, is expressed through the decline in passenger ships in 2020 in Figure 5.5b and 5.8. Ultimately, the time frame in this thesis is limited and does not capture the yearly changes in climatic variables. For example, Figure 5.1b neither shows increased SST nor reduced SIC by time, which otherwise has been proven in the literature. This may further explain the low importance of climatic variations and the model predictions.

Despite the limitations of the time frame, it seems plausible that monthly variations influence activity according to the Arctic navigation season. Figure 5.1b and the feature importances of Case Study 1 in Figure 5.4 indicate a poor correlation between activity presence and the *month*-encoded feature. This is potentially caused by the cargo ships, which constitute the highest proportion of unique ships in the overall data, but have no obvious seasonal variations, according to Figure 5.6b and the low feature importance in Figure 5.11b. On the other hand, the fishing vessels and cruise ships correlate better with the seasonal variations. These differences can be explained by Figure 5.8 which shows that cargo ships, on average, remain at the lower latitudes throughout the years. In contrast, the fishing vessels and the passenger ships operate further north according to the navigation access during summer and the seasonal demand in fishing areas and tourism, respectively.

6.4 Applicability of the proposed solution

This thesis attempts to establish an ML-aided basis for estimating activity presence and vessel densities by industry from independent factors describing the Arctic surroundings. The fundamental motivation is that increased insight into activity patterns will contribute to understanding the Arctic risk picture that follows from an increased number of vessels operating in the region. Although this thesis investigates the possibility of using ML for activity prediction and how different activity targets correlate with independently extracted predictors, the next step, in terms of risk prevention, would be to apply similar predictors for estimating *future* activity.

Whereas spatial coordinates, ocean depth, and distance measures are independent of time, climatic variations such as ice and temperature evolve by both time and space. However, as mentioned in Section 2.4, the future conditions of such variables are easily retrieved from climate models. Consequently, these features may be transferred to scenarios considering other spatial regions and even time frames not yet observed. Nevertheless, the main obstacle in further applying the proposed solution relies on the sudden events that are not as easily transferable to ML features. However, they may have significant impacts on the actual activity measure. These events include economic and political shifts, evolved infrastructure, restrictions, regulations, and fish fluctuations. Covid-19 is another example from which consequences in tourism activities have already been mentioned. As elaborated in Section 2.3, the Arctic is not governed by one single regime. Therefore, it is reasonable to assume that such events frequently will appear in the future of an open Arctic.

In the proposed solution, the *year* feature acts as a simple proxy that measures up these multiple, sudden impacts on activity. Whereas year itself does not necessarily change activity, there exist a relationship derived by the disruptive circumstances that occur *within* each year. However, as this underlying information is neither provided explicitly to the model nor follows a specific trend over time, patterns will be missed according to the sudden changes that the ML model cannot foresee and control. This further substantiates the activity targets' obscure dependencies with time.

Chapter 7

Conclusions and Further Work

This master thesis has targeted the feasibility of applying supervised learning methods for estimating activity and evaluating sources of activity trends in a part of the Barents Sea within the Arctic region. Based on data coverage, the selected period is from January 2015 to May 2021. The work was driven by three research questions, which were investigated through systematic background studies, data allocation, data processing, model implementations, and experimental tests by two case studies. This chapter concludes the thesis in light of the research questions and suggests further work.

7.1 Conclusions

The first research question addressed was *"How to develop a supervised ML framework for predicting vessel presence and vessel density in the Arctic waters by time and space?"* The proposed solution was described by an ML workflow covering the steps of ML training data development and algorithm implementation. Explanatory predictors describing the Arctic environment and climate were collected from multiple sources. The data sets were individually processed according to data formats, missing values, irrelevance, and erroneous values detected on land. However, they were provided in different resolutions by time and space. Therefore, a spatiotemporal grid was created to manage assembling the data into a composed ML training data set. The grid cells were used as key variables for data aggregation and merging. In addition, ASTD AIS data were used to extract the number of unique vessels within each grid cell, which formed the basis for the ML prediction target.

For performance review of the proposed solution and validation of the remaining research questions, two case studies were defined from different viewpoints of activity. Whereas Case

Study 1 investigated the detection of vessel presence within the spatiotemporal region, Case Study 2 was restricted to the case of vessel presence only and considered estimation of vessel density within fishing, cargo shipping, and tourism. Based on background from the literature, the supervised ensemble models RF and XGBoost were selected and applied to the case studies with functionality support from their respective Python libraries. Although the data suffered from imbalanced target distributions, the analyses of the results proved that the classifiers performed with averaged F1-scores of around 80% in both case studies. The binary task of Case Study 1 obtained the best accuracy and F1-scores, by 90% and 87%, respectively. The recall and precision values did not diverge significantly, which further substantiates the satisfactory performance. The classifiers' relatively similar results and learning behavior justifies their individual performance.

The second research question addressed was *"How do ecological and environmental factors in the Arctic affect activity presence and the level of vessel density?"* The ecological and environmental factors were represented by the predictors of the training data, from which impact on prediction was verified by investigating their correlation with the target values and assessing which predictors the classifiers applied the most during model training. The most significant predictor was *latitude*, which negatively correlated with both activity presence and vessel density. In addition, SST proved to have a high impact on classification in Case Study 1, which verified the expectation of more activity presence in warmer temperatures. Such a temperature influence was not observed in Case Study 2 due to its concentration at the lower latitudes where the vessel densities are primarily present and the temperature changes according to the polar front are less distinct. Similarly, the SIC attribute did not contribute significantly to prediction, yet, no yearly reduction in sea ice was identified for the given period. Whereas the fishing and tourism industries proved to extend toward the north during the summer season, a general northern activity increase by year was not observed. Moreover, cargo ships behaved independently of the season but showed a clear connection with closer distances to the coast. The latter also applied to the fishing vessel densities.

The final research question addressed was *"Is supervised ML a reliable approach for activity predictions in the Arctic?"* Despite irregular target distributions, the overall classifiers predicted by relatively high scores in both recall and precision, which justifies the power of ML to recognize activity patterns in vast amounts of unstructured data. However, deviations were observed between the training and the validation processes from model learning, meaning that the provided data were insufficient for confident predictions. In addition, the model relied mainly on the latitude feature, which caused erroneous predictions of classes where the spatial

separation was less clear. Ultimately, the model is limited to only learning patterns from what information it is provided and is unable to foresee disruptive circumstances such as rules, regulations, and political shifts that most certainly will affect the future Arctic. Consequently, there remains a challenge in explicitly encapsulating all influential factors affecting activity in one single ML model and applying it for long future time scales. Therefore, the results should not be fully trusted but viewed as indicators of activity.

7.2 Further work

There are various measures that can be applied in order to increase the reliability and applicability of the proposed supervised learning solution. Firstly, the *Unknown* ASTD categorized vessels could be imputed by using deep learning for pattern recognition of the vessels' trajectories. Consequently, the deviations from the actual amount of vessel activity represented in the Arctic will be reduced, and more realistic activity estimations will follow.

Secondly, as real-world data always will involve imbalance challenges, alternative resampling approaches should be investigated. The SMOTETomek algorithm applied in this thesis follows the nearest neighbor approach. However, there exist various validated candidate approaches that are based on other clustering methods.

Thirdly, individual investigations of the predictors' impact on model performance could be conducted, such as the investigation of model performance by excluding the *latitude* feature. Then, it should be considered to apply additional features to the training data to contribute explicit information toward activity estimation. Such features would include industry-specific features describing supply and demand, as well as spatiotemporal climatic changes, including humidity, windiness, fogginess, and atmospheric pressure, which are expected to fluctuate and affect the Arctic region.

Finally, a better impression of how activities affect the pristine Arctic could be obtained by transforming the ML model to a region not affected by operations close to the mainland coastline. In addition, more years could be included to capture the climatic changes appropriately. However, the application of the model for future activity estimations should be performed carefully and over a short time scale due to the possibility of sudden political or regulatory changes that the model cannot foresee.

Bibliography

Arctic Council (2009). *Arctic Marine Shipping Assessment 2009 Report*. Arctic Council.

– (June 2021). *An Introduction to: the International Agreement to Prevent Unregulated Fishing in the High Seas of the Central Arctic Ocean*. URL: https://www.arctic-council.org/news/introduction-to-international-agreement-to-prevent-unregulated-fishing-in-the-high-seas-of-the-central-arctic-ocean/. (Accessed: 08.03.2022).

Arctic Portal (n.d.). *Arctic Council*. URL: https://arcticportal.org/arctic-governance/arctic-council. (Accessed: 31.01.2022).

ASTD PAME (Apr. 2020). *Arctic Ship Traffic Data User Guide*. Arctic Council.

– (Mar. 2021). *ASTD Data*. Arctic Council.

AWI (May 2020). *Arctic Governance*. Alfred-Wegener-Institut.

Azzara, A. J., H. Wang, and Daniel Rutherford (Sept. 2019). *A Ten-Year Projection of Maritime Activity in the U.S. Arctic Region, 2020–2030*.

Banarjee, P. (2020). *A Guide on XGBoost hyperparameters tuning*. URL: https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning/notebook. (Accessed: 01.5.2022).

Batista, Gustavo, Ronaldo Prati, and Maria-Carolina Monard (June 2004). "A Study of the Behavior of Several Methods for Balancing machine Learning Training Data". In: *SIGKDD Explorations* 6, pp. 20–29.

Benz, L., C. Münch, and E. Hartmann (Aug. 2021). "Development of a search and rescue framework for maritime freight shipping in the Arctic". In: *Transportation research* 152, pp. 54–69.

Breiman, L. (Oct. 2001). "Random Forests". In: *Machine Learning* 45, pp. 5–32.

British Antarctic Survey (Apr. 2020). *Report on Geospatial Analysis of Arctic Marine Tourism - Phase 1*. Arctic Council.

Brownlee, Jason (2020). *Feature Importance and Feature Selection With XGBoost in Python*. URL: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/. (Accessed: 6.12.2021).

Caruana, R. and A. Niculescu-Mizil (2006). "An Empirical Comparison of Supervised Learning Algorithms". In: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: Association for Computing Machinery, pp. 161–168.

Chatzikokolakis, K. et al. (May 2019). "A comparison of supervised learning schemes for the detection of search and rescue (SAR) vessel patterns". In: *GeoInformatica* 25.

Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chen, T. and C. Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: ACM, pp. 785–794.

Copernicus (2021). DOI: https://doi.org/10.48670/moi-00123. (Accessed: 26.04.2022).

Cracknell, M. J. and A. M. Reading (2014). "Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information". In: *Computers & Geosciences* 63, pp. 22–33.

Davis, John C and Robert J Sampson (1986). *Statistics and data analysis in geology*. Vol. 646. Wiley New York.

EPPR, Arctic Council (2018). *Guideline for Arctic Marine Risk Assessment*. URL: https://eppr.dnvgl.com/. (Accessed: 04.11.2021).

Fauchald, Per et al. (2021). "Poleward shifts in marine fisheries under Arctic warming". In: *Environmental Research Letters* 16.7.

Giffen, Benjamin van, Dennis Herhausen, and Tobias Fahse (2022). "Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods". In: *Journal of Business Research* 144, pp. 93–106.

Global Fishing Watch (n.d.). *What vessels are required to use AIS? What are global regulations and requirements for vessels to carry AIS?* URL: https://globalfishingwatch.org/faqs/what-vessels-are-required-to-use-ais-what-are-global-regulations-and-requirements-for-vessels-to-carry-ais/. (Accessed: 19.11.2021).

Goel, Garima et al. (2013). "Evaluation of sampling methods for learning from imbalanced data". In: *International Conference on Intelligent Computing*. Springer, pp. 392–401.

Goerlandt, F. et al. (Feb. 2017). "An analysis of wintertime navigational accidents in the Northern Baltic Sea". In: *Safety Science* 92, pp. 66–84.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Gridded Bathymetry Data - Arctic Ocean (IBCAO) (2020). URL: https://www.gebco.net/data_and_products/gridded_bathymetry_data/arctic_ocean/. (Accessed: 01.02.2022).

Harris, C. R. et al. (Sept. 2020). "Array programming with NumPy". In: *Nature* 585, pp. 357–362.

Hollowed, A. B., B. Planque, and H. Loeng (2013). "Potential movement of fish and shellfish stocks from the sub-Arctic to the Arctic Ocean". In: *Fisheries Oceanography* 22.5, pp. 355–370.

Høyer, J., W. M. Kolbe, et al. (Nov. 2021). *Product User Manual*. Copernicus Marine Service.

Høyer, J., M. H. Riebergaard, et al. (Dec. 2021). *Quality Information Document*. Copernicus Marine Service.

Høyer, J. and J. She (Mar. 2007). "Optimal interpolation of sea surface temperature for the North Sea and Baltic Sea". In: *Journal of Marine Systems* 65, pp. 176–189.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95.

ICES (2018). *Barents capelin bounce back – and other highlights*. URL: https://www.ices.dk/news-and-events/news-archive/news/Pages/Barents-capelin-bounce-back---and-other-highlights.aspx. (Accessed: 25.5.2022).

– (Mar. 2022). *Barents Sea Ecoregion – Fisheries overview*.

IMO (2007). "Consolidated text of the guidelines for Formal Safety Assessment (FSA) for use in the IMO rule-making process". In: *MSC/Circ* 1023.

– (2017). *International Code for Ships Operating in Polar Waters (Polar Code)*. URL: https://www.imo.org/en/OurWork/Safety/Pages/polar-code.aspx. (Accessed: 10.02.2022).

– (n.d.). *AIS transponders*. URL: https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx. (Accessed: 31.01.2022).

Jakobsson, M., L. Mayer, et al. (July 2020). "The International Bathymetric Chart of the Arctic Ocean Version 4.0". In: *Scientific Data* 7, p. 14.

Jakobsson, M., L.A. Mayer, and C. Bringensparr (July 2020). "The International Bathymetric Chart of the Arctic Ocean Version 4.0". In: *Sci Data* 7.

Jensen, L. and E. Paglia (Apr. 2021). *Assessing the future of Arctic shipping in the wake of the Suez Canal incident*. URL: https://open.spotify.com/episode/1V8SfbjErlAF2jVonS2BKP?si=b0ba37074fb44904&nd=1.

Jiang, X. et al. (May 2016). "Fishing Activity Detection from AIS Data Using Autoencoders". In: pp. 33–39.

Johannesen, Edda et al. (2021). "Fish diversity data from the Barents Sea Ecosystem Survey 2004-2019". In: *Rapport fra havforskningen*.

Khalitov, Ruslan (Oct. 2021a). *Ensemble Learning, TDT4173 Machine Learning*.

– (Oct. 2021b). *ML in Practice, TDT4173 Machine Learning*.

Kluyver, T. et al. (Dec. 2016). *Jupyter Notebooks - a publishing format for reproducible computational workflows*.

Kotsiantis, Sotiris, D. Kanellopoulos, and P. Pintelas (Nov. 2005). "Handling imbalanced datasets: A review". In: *GESTS International Transactions on Computer Science and Engineering* 30, pp. 25–36.

Kotu, Vijay and Bala Deshpande (2015). "Chapter 8 - Model Evaluation". In: *Predictive Analytics and Data Mining*. Morgan Kaufmann, pp. 257–273.

Kraus, P., C. Mohrdieck, and F. Schwenker (2018). "Ship classification based on trajectory data with machine-learning methods". In: *2018 19th International Radar Symposium (IRS)*, pp. 1–10.

Kuhn, M. and K. Johnson (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. 1st. Chapman and Hall/CRC.

Lahn, Glada and Charles Emmerson (2012). *Arctic opening: Opportunity and risk in the high north*.

Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas (2017). "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17, pp. 1–5.

Li, Huanhuan et al. (2017). "A dimensionality reduction-based multi-step clustering method for robust vessel trajectory analysis". In: *Sensors* 17.8.

Liu, M. and J Kronbak (2010). "The potential economic viability of using the Northern Sea Route (NSR) as an alternative route between Asia and Europe". In: *Journal of Transport Geography* 18.3, pp. 434–444.

Liu, M., M. Wang, et al. (2013). "Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar". In: *Sensors and Actuators B: Chemical* 177, pp. 970–980.

Liu, Xian (2016). "Chapter 14 - Methods for handling missing data". In: *Methods and Applications of Longitudinal Data Analysis*. Ed. by Xian Liu. Oxford: Academic Press, pp. 441–473.

Loe, J et al. (2014). "Arctic business scenarios 2020. Oil in demand, green transformation, re-freeze". In: *Menon Business Economics* 41.

Marchenko, N. A. (2019). “"Marine Emergencies in the Arctic" - GIS Online Resource for Preparedness, Response and Education”. In: *Port and Ocean Engineering under Arctic Conditions Conference*.

Marchenko, N. A. et al. (Mar. 2018). “Arctic Shipping and Risks: Emergency Categories and Response Capacities”. In: *The International Journal on Marine Navigation and Safety of Sea Transoportation* 12.1.

Marsh Risk Management Research (Aug. 2014). *ARCTIC SHIPPING: NAVIGATING THE RISKS AND OPPORTUNITIES*. Marsh Risk Management Research.

Mazzarella, F. et al. (July 2014). “Discovering vessel activities at sea using AIS data: Mapping of fishing footprints”. In: *FUSION 2014 - 17th International Conference on Information Fusion*.

McKinney, Wes (2010). “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 56–61.

MIT (May 2021). *Collaborative Data Science for Healthcare*. URL: https://openlearninglibrary.mit.edu/courses/course-v1:MITx+HST.953x+3T2020/course/.

Mitchell, T. M. (1997). *Machine Leaning*. 1st. McGraw-Hill Science/Engineering/Math.

Mohit (2020). *Random Forest Hyperparameter tuning*. URL: https://www.kaggle.com/code/mohitsital/random-forest-hyperparameter-tuning/notebook. (Accessed: 01.5.2022).

NASA Ocean Biology Processing Group (OBPG) (2012). URL: http://pacioos.org/metadata/dist2coast_1deg.html. (Accessed: 01.02.2022).

NATO Shipping Centre (2021). *AIS (AUTOMATIC IDENTIFICATION SYSTEM) OVERVIEW*. URL: https://shipping.nato.int/nsc/operations/news/2021/ais-automatic-identification-system-overview. (Accessed: 24.04.2022).

NCAS (n.d.). *What is a climate model?* URL: https://ncas.ac.uk/learn/what-is-a-climate-model/. (Accessed: 17.02.2022).

netCDF4 (Feb. 2015). URL: http://unidata.github.io/netcdf4-python/.

Nguyen, Duong et al. (2018). “A multi-task deep learning architecture for maritime surveillance using AIS data streams”. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 331–340.

Norwegian Polar Institute (n.d.). *Barents Sea*. URL: https://www.npolar.no/en/themes/barents-sea/. (Accessed: 08.03.2022).

NSIDC (2020). *What is the Arctic?* URL: https://nsidc.org/cryosphere/arctic-meteorology/arctic.html. (Accessed: 21.02.2022).

Ocean Conservancy (2017). *Navigating the North: An Assessment of the Environmental Risks of Arctic Vessel Traffic*. Ocean Conservancy.

Øien, K. (June 2000). "Risk indicators as a tool for risk control". In: *Reliability Engineering & System Safety* 74, pp. 129–145.

PAME (May 2021a). *Arctic Marine Tourism Project Report*. Arctic Council.

– (Mar. 2021b). *Arctic Shipping Status Report (ASSR) #1*. Arctic Council.

– (Apr. 2021c). *Arctic Shipping Status Report (ASSR) #3*. Arctic Council.

– (May 2021d). *ASTD Data*. PAME, Arctic Council.

PAME - Arctic Ship Traffic Data (2022). URL: http://ASTD.is. (Accessed: 01.02.2022).

Paxian, A. et al. (2010). "Present-Day and Future Global Bottom-Up Ship Emission Inventories Including Polar Routes". In: *Environmental Science & Technology* 44.4, pp. 1333–1339.

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

Peters, G. P. et al. (2011). "Future emissions from shipping and petroleum activities in the Arctic". In: *Atmospheric Chemistry and Physics* 11.11, pp. 5305–5320.

PyProj (2022). URL: https://pyproj4.github.io/pyproj/v2.2.2rel/index.html.

Russel, S. and P. Norvig (2020a). *Artificial Intelligence: A Modern Approach*. 3rd. Pearson.

– (2020b). *Artificial Intelligence: A Modern Approach*. 3rd. Pearson.

Rusten, M. et al. (Mar. 2015). "ARCTIC RISK PICTURE – MANAGEMENT OF SAFETY AND ENVIRONMENTAL RISK". In: 12th Offshore Mediterranean Conference.

FN-sambandet (n.d.). *Den internasjonale sjøfartsorganisasjonen (IMO)*. URL: https://www.fn.no/om-fn/fns-organisasjoner-fond-og-programmer/den-internasjonale-sjoefartsorganisasjonen-imo. (Accessed: 26.04.2022).

Sarker, Iqbal H (2021). "Machine learning: Algorithms, real-world applications and research directions". In: *SN Computer Science* 2.3, pp. 1–21.

Sasada, Taisho et al. (2020). "A Resampling Method for Imbalanced Datasets Considering Noise and Overlap". In: *Procedia Computer Science* 176, pp. 420–429.

Sea Around Us (June 2015). *Sea Around Us Area Parameters and Definitions*. URL: https://www.seaaroundus.org/sea-around-us-area-parameters-and-definitions/#_Toc421807917. (Accessed: 03.03.2022).

Shevelev, M. and H. Gjøsæter (June 1999). *Overview of Fish Stocks in the Barents Sea and Asjacent Areas*.

Stenseth, Nils Chr. (n.d.). *NMDC: Norwegian Marine Data Centre*. URL: https://www.mn.uio.no/cees/english/research/projects/190608/. (Accessed: 14.05.2022).

Stephenson, S. R., L. C. Smith, and J. Agnew (May 2011). "Divergent long-term trajectories of human access to the Arctic". In: *Nature Climate Change* 1, pp. 156–160.

Stephenson, S. R., L. C. Smith, and L. W. Brigham (June 2013). "Projected 21st-century changes to Arctic marine access". In: *Climate Change* 118, pp. 885–899.

Svalbard Museum (n.d.). *THE SEA*. URL: https://svalbardmuseum.no/en/natur/havet/. (Accessed: 12.12.2021).

The Arctic Institute (Sept. 2015). *The Future of the Northern Sea Route – A "Golden Waterway" or a Niche Trade Route*. URL: https://www.thearcticinstitute.org/future-northern-sea-route-golden-waterway-niche/. (Accessed: 25.05.2022).

Tomek, Ivan (1976). *Two Modifications of CNN*.

TRAFICOM (Jan. 2019). *Guidelines for the Application of the 2017 Finnish-Swedish Ice Class Rules*. Finnish Transport and Communications Agency.

Tukey, J. M. (1977). *Exploratory Data Analysis*. 1st. Reading, Mass.

United Nations (May 2017). *The Role of the International Maritime Organization in Preventing the Pollution of the World's Oceans from Ships and Shipping*. URL: https://www.un.org/en/chronicle/article/role-international-maritime-organization-preventing-pollution-worlds-oceans-ships-and-shipping. (Accessed: 26.04.2022).

Van Rossum, G. and F. L. Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. ISBN: 1441412697.

Virtanen, P. et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272.

Wang, Zhe et al. (2019). "SMOTETomek-Based Resampling for Personality Recognition". In: *IEEE Access* 7.

Wessel, Paul et al. (2019). "The generic mapping tools version 6". In: *Geochemistry, Geophysics, Geosystems* 20.11, pp. 5556–5564.

Whiteman, G. et al. (2021). *WHAT ARCTIC BREAKDOWN MEANS FOR COP26*. Arctic Basecamp.

Wolsing, Konrad et al. (2022). "Anomaly Detection in Maritime AIS Tracks: A Review of Recent Approaches". In: *Journal of Marine Science and Engineering* 10.1, p. 112.

Yellenki, Ritheesh Baradwaj (2020). *Top 8 Challenges for Machine Learning Practitioners*. URL: https://towardsdatascience.com/top-8-challenges-for-machine-learning-practitioners-c4c0130701a1. (Accessed: 24.04.2022).

Zhong, H., X. Song, and L. Yang (2019). "Vessel Classification from Space-based AIS Data Using Random Forest". In: *2019 5th International Conference on Big Data and Information Analytics (BigDIA)*, pp. 9–12.

Appendices

A Data sources

The table below presents an overview of all data sources investigated as potential candidates for constructing the ML data sets of predictors and targets.

Source	Description	Format(s)
Vessel traffic data		
PAME – Arctic Ship Traffic Data	Historical data of ship tracks, ship types, ports, ship characteristics, emissions and fuel consumption in the Arctic. Restricted access, permission required. Provided by ASTD PAME.	Tabular (**CSV**)
Norwegian Directorate of Fisheries	Norwegian fishing vessel VMS data from 2011 until the present. Includes ship tracks, ship characteristics and ownership information.	Tabular (**CSV**)
Norwegian Coastal Administration (Kystdatahuset)	Data covering vessel activities and traffic around the Norwegian coast. Ability to download AIS samples from small time ranges. Account required. Provided by The Norwegian Coastal Administration.	Tabular (**CSV**)
Barents Watch	AIS data from the Norwegian coast and sea areas. Data from the past 24 hours. Account required.	Tabular (**CSV**)
Global Fishing Watch	Data covering global fishing activity. Account required.	Tabular (**CSV**)
Metocean and ecological data		
Svalbard Integrated Arctic Earth Observing System	Data from long-term measurements in/ around the Norwegian archipelago of Svalbard. Earth System Science in Svalbard. Data on bathymetry, atmospherical, environmental conditions at sea surface. Some data with restricted access.	Raster, Tabular (**NetCDF**, **CSV**)
International Bathymetric Chart of the Arctic Ocean	Digital database of bathymetric data north of 64 degrees.	Raster (**NetCDF**)
NASA's Ocean Biology Processing Group	Provide collected and processed satellite-based information of ocean biology and climate-related inquiries.	Raster (**NetCDF**)
Barents Watch	Digital information service, ArcticInfo. Spatial data describing ice concentration, icebergs and ice edges. Account required.	Raster (**Shape**)
Polar View	Sea ice charts and concentration. Provided by U.S. National Ice Centre and AMSR2 from University of Bremen.	Raster (**GeoTiff, ESRI**)

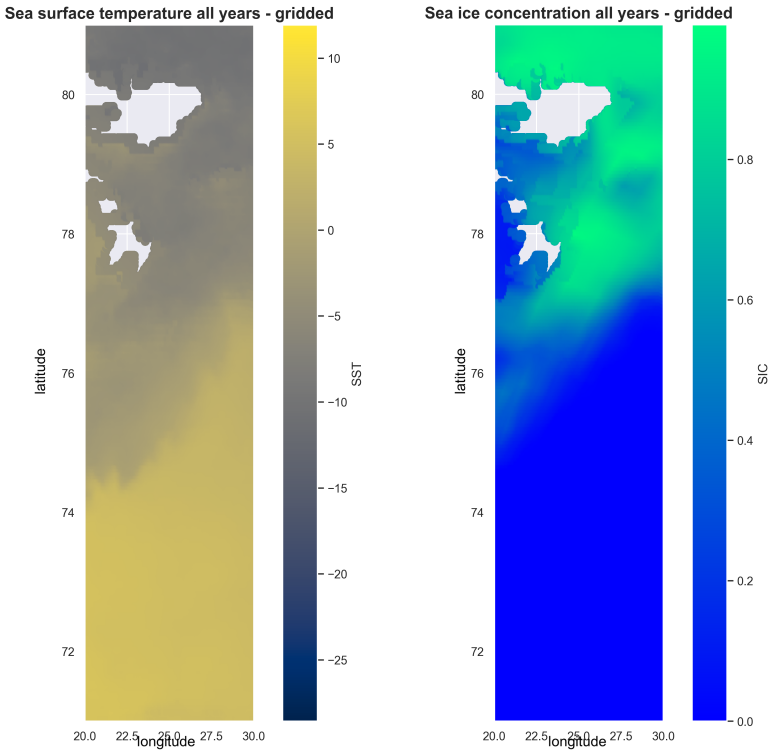
Norwegian Maritime Data Centre	Maritime data covering sea areas important for Norway.	Raster, tabular (**Shape, CSV**)
Norwegian Polar Data Centre	Various of ecological and physical data. Poor filtering mechanism.	Raster, tabular (**Shape, CSV**)
MAREANO	Data describing depth, sea bed conditions, biodiversity and pollution at sea.	Photography (**tiff**)
Ocean Biodiversity Information System	Open-access data providing global information on marine biodiversity. Pulls together data from sources around the world. Spatial and temporal gaps.	Tabular (**CSV**)
Global Biodiversity Information Facility	Open-access data on biodiversity and wildlife. Account required.	Not applicable
National Snow and Ice Data Center	Scientific data for research on snow, glaciers, sea ice, frozen ground etc. Easy filtering mechanism. Access to NASA Earthdata required for some of the data sets.	Raster, Tabular (**NetCDF**, **Shape**, **binary**, **GeoTiff**, **CSV**)
General Bathymetric Chart of the Oceans	Gridded bathymetry data sets for the world' s oceans.	Raster (**NetCDF**, **GeoTiff**)
Copernicus Marine Service	Open access marine data, physical and environmental. Account required.	Raster (**NetCDF**)
Norwegian Coastal Administration (Kystdatahuset)	Data on Norwegian ports, navigation and wave forcasts. Provided by The Norwegian Coastal Administration and Geonorge.	Raster (**Shape**)

Table A.1: Data sources investigated during data allocation

B EDA output

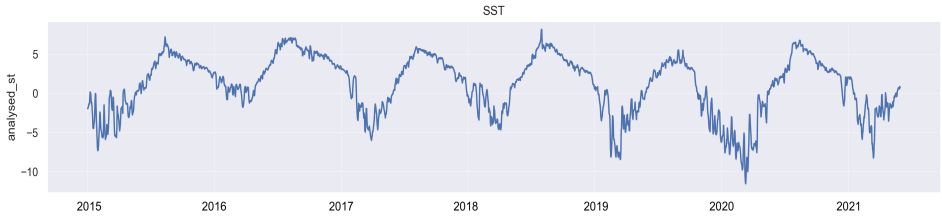
The plots and graphs provided below are a subset of the output carried out from the EDA. In addition to the other output presented previously, the plots were used to gain insight into the data in terms of data quality, errors, attribute connections, and coverage. Most of them are related to the ASTD AIS data set, which involves several attributes.

Copernicus

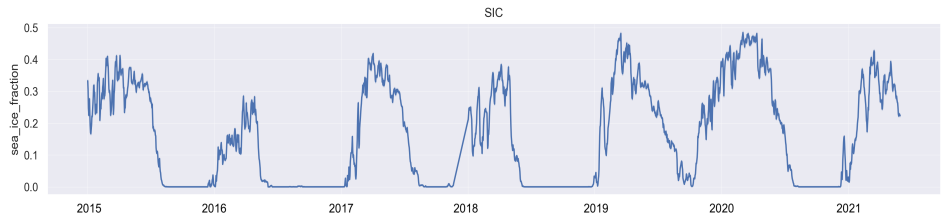


(a) SST (b) SIC

Figure B.1: Spatial plots of gridded Copernicus data within the whole time frame. A connection between lower temperatures and higher sea ice concentrations is observed according to the polar front around latitude 76°



(a) Mean SST evolution. Higher temperatures during summer time (middle of each year)



(b) Mean SIC evolution. Higher sea ice concentrations during winter time

Figure B.2: Temporal plots of Copernicus data (spatial mean). A comparison of the two plots indicates a negative correlation by season

ASTD - All Records

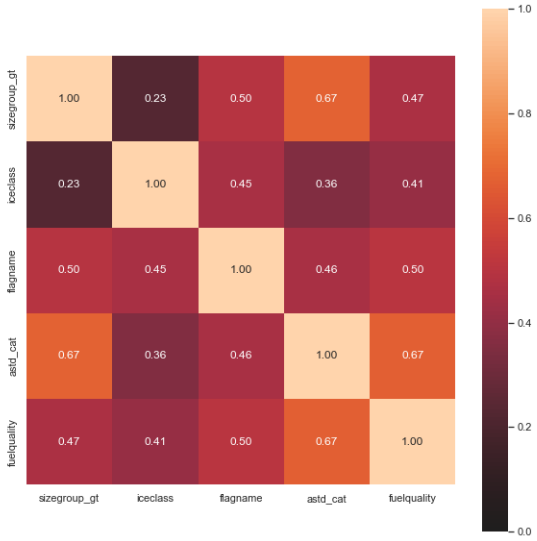


Figure B.3: Heat map of Pearson product-moment coefficients between the static attributes. Only positive correlations are observed. Size group and ASTD category have the highest correlation

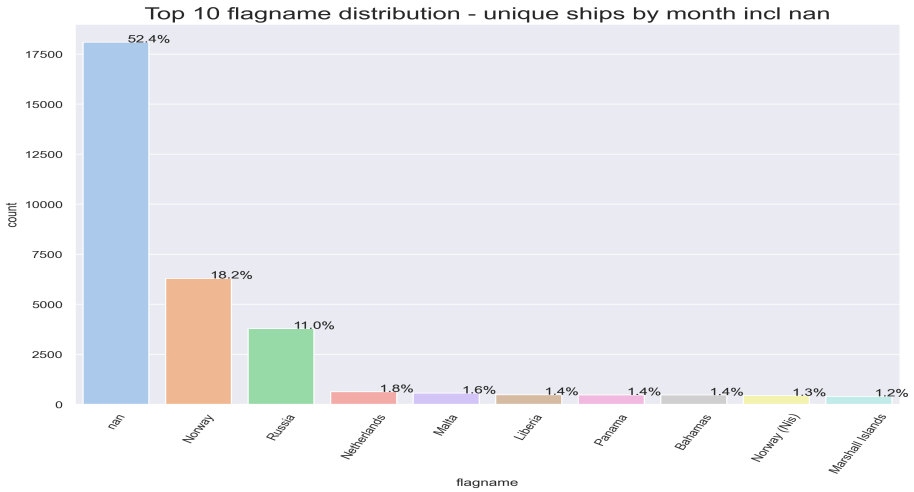


Figure B.4: Distribution in flag names. Most flag name values are NaN-categorized. Norway and Russia are the most represented countries

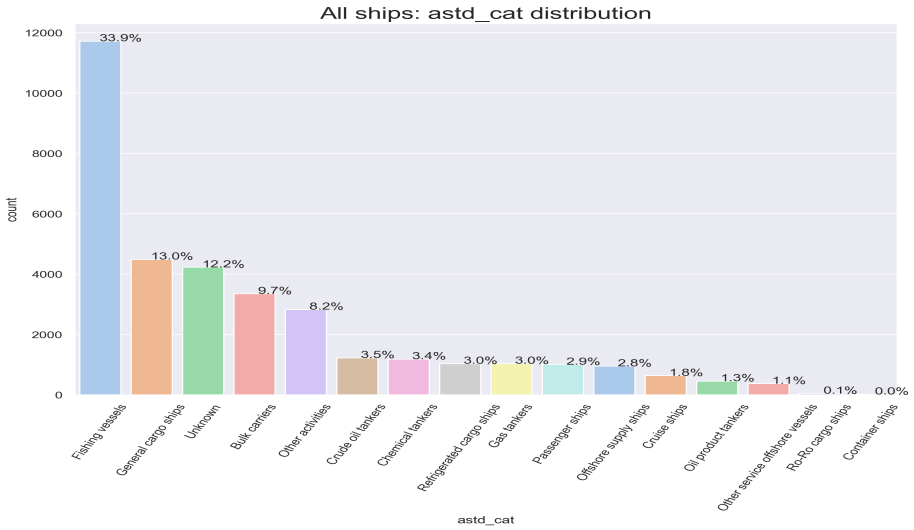


Figure B.5: Distribution in original ASTD categories (by unique vessels). Fishing vessels have the highest share

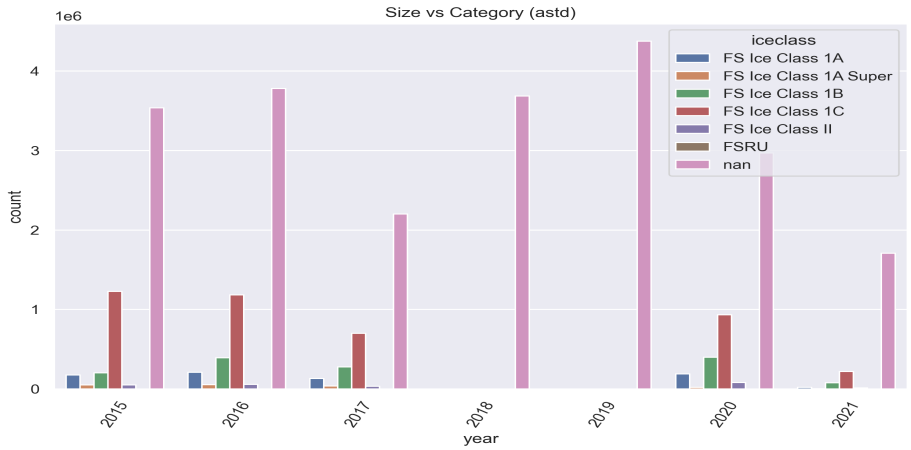


Figure B.6: Distribution in registered ice classes. Most ice class values are missing, including all values of 2018 and 2019

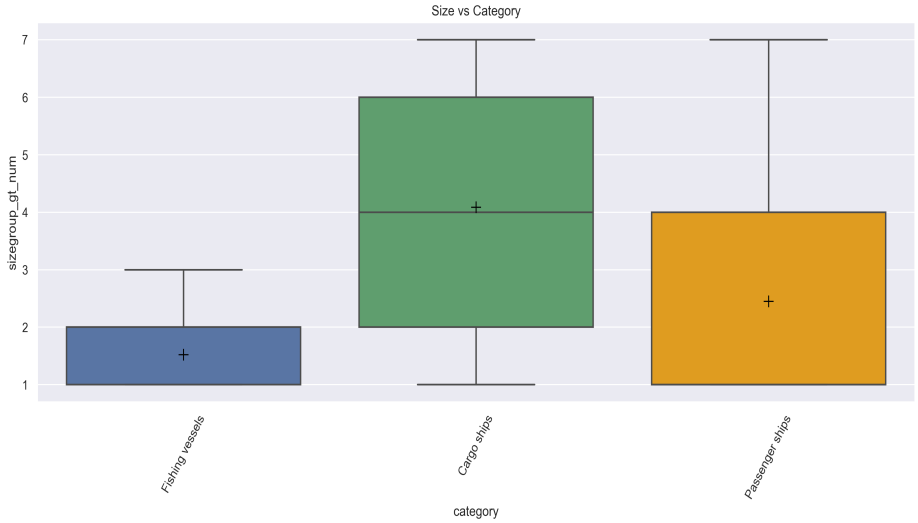


Figure B.7: Box plot combining size and overall category (as defined in this thesis). The higher the size group number, the larger the vessel. Fishing vessels constitute smaller size groups. Cargo ships and passenger ships span all sizes according to different ASTD categories

ASTD - Fishing vessels

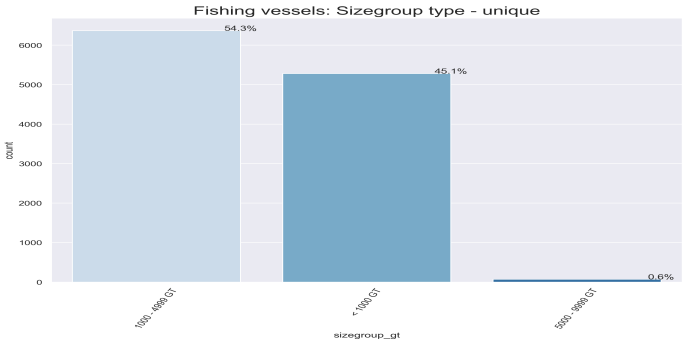
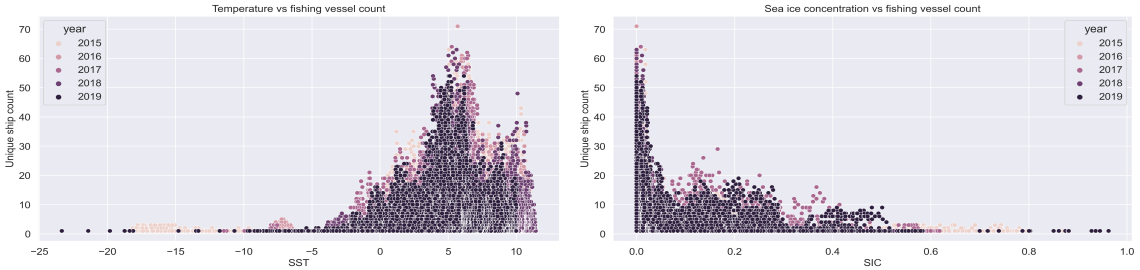


Figure B.8: Distribution in size group among fishing vessels (number of samples). Most fishing vessels are small

(a) Unique ship count vs. bathymetry (km) and year: no clear tendency

(b) Mean distance from vessels to coast vs. time: most distances are above 70 km. Slight increase from 2015 to 2020

Figure B.9: Fishing vessels vs. spatial features

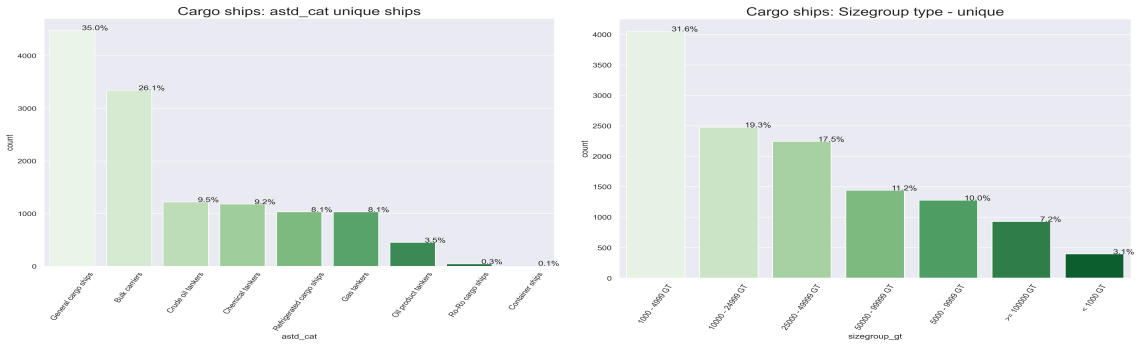


(a) SST: increase in activity with higher temperatures. Looks like there was more activity for colder temperatures in 2015 compared to the present. Higher activities around 5 degrees

(b) SIC: the highest densities are observed where there is no sea ice concentration. 2015 involves several cases of high sea ice concentrations (above 0.6)

Figure B.10: Fishing vessels vs. climatic features

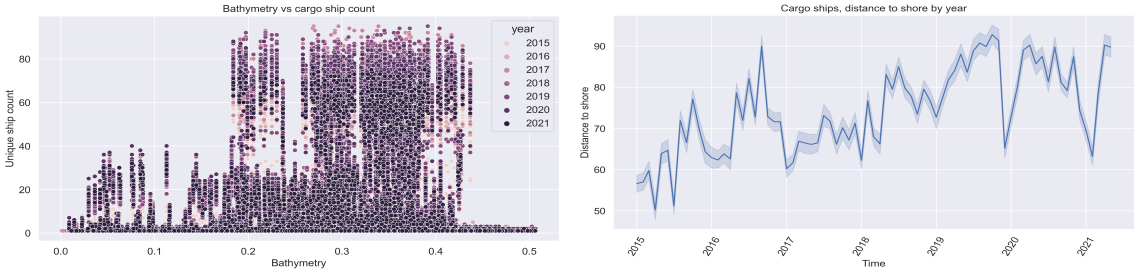
ASTD - Cargo ships



(a) ASTD category distribution. Numerous ASTD cargos/tankers/carriers included. General cargo ships have the highest share

(b) Size group distribution. The size range covers all sizes according to the many ASTD categories represented. The second smallest group stands out

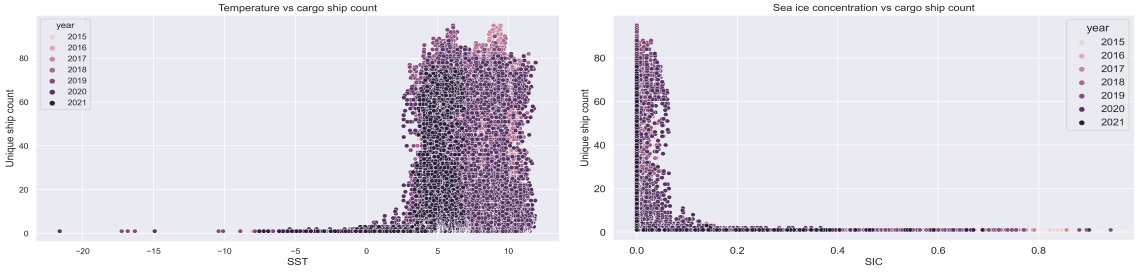
Figure B.11: Cargo ships distributions by count plots (number of samples)



(a) Unique ship count vs. bathymetry (km) and year: there is a higher share of unique cargo ships where the ocean is deep

(b) Mean distance from ships to coast vs. time: there is an increasing tendency for larger distances by year. Sudden peaks to smaller distances at year-end in recent years

Figure B.12: Cargo ships vs. spatial features

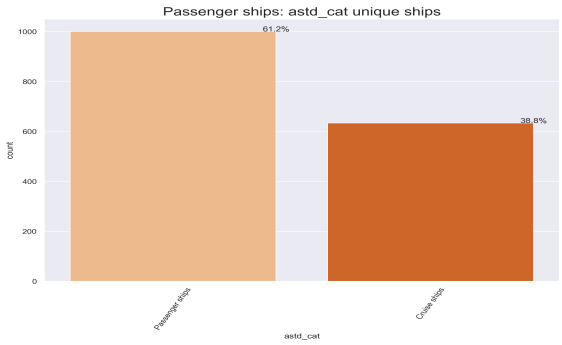


(a) SST: most cargo ships operate when the temperature is above 0°C. No specific pattern by year is observed

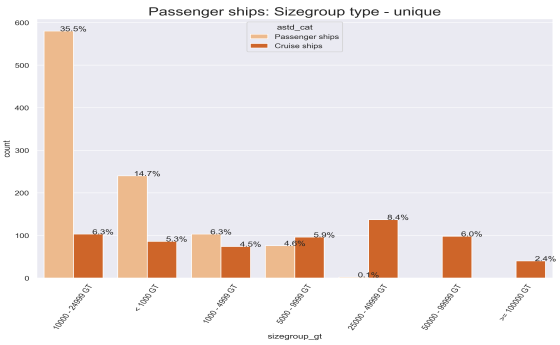
(b) the highest densities are observed where there is no sea ice concentration. However, no clear correlation is observed

Figure B.13: Cargo ships vs. climatic features

ASTD - Passenger ships

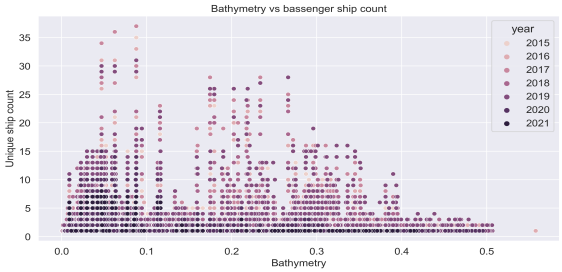


(a) ASTD category distribution. Both general passenger ships and cruise ships are well represented

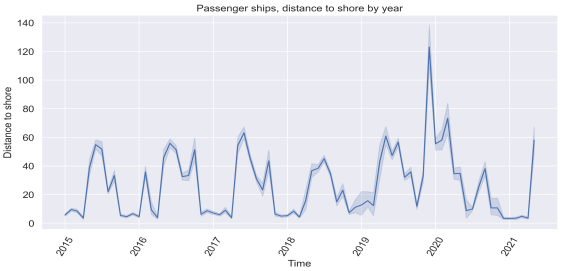


(b) Size group distribution. Cruise ships are represented among all groups and constitute the largest ships. Most passenger ships are within the medium size group

Figure B.14: Passenger ships distributions by count plots (number of samples)

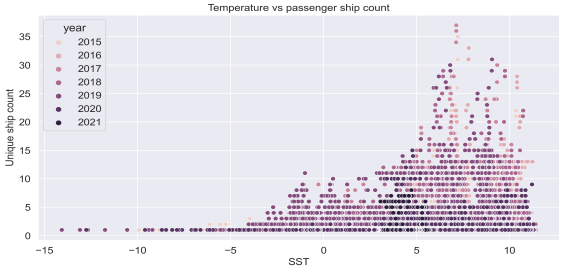


(a) Unique ship count vs. bathymetry (km) and year: more activity in shallow water, but no clear tendency by year

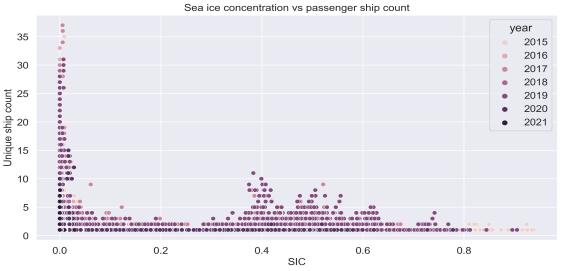


(b) Mean distance from ships to coast vs. time: distance increases by season. There is a notable peak in the distance at the year-end of 2020

Figure B.15: Fishing vessels vs. spatial features

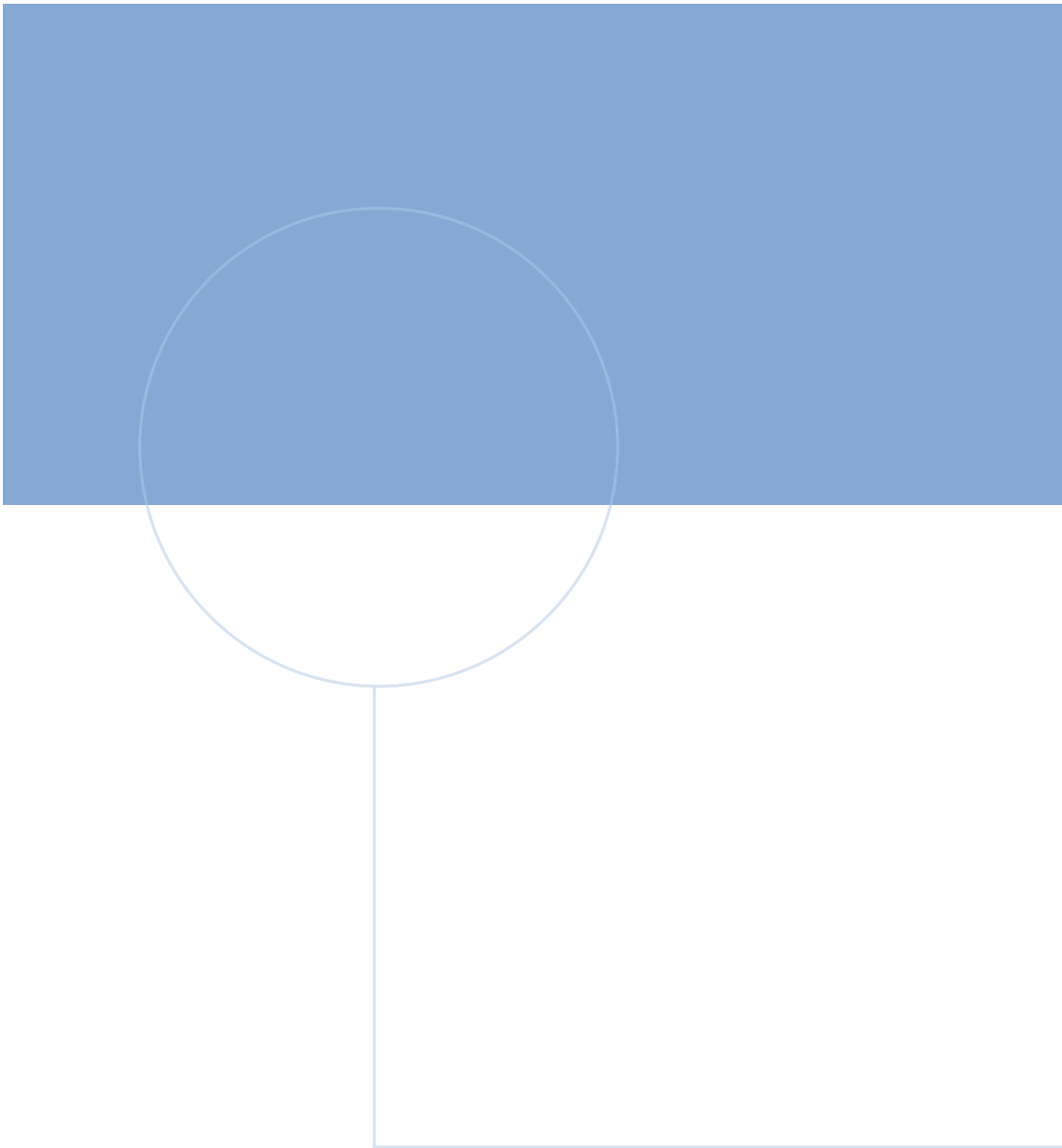


(a) SST: increase in activity by higher temperatures



(b) SIC: the highest densities are observed where there is no sea ice fraction. Higher concentration in 2015 compared to other years (above 0.8)

Figure B.16: Fishing vessels vs. climatic features



NTNU
Norwegian University of
Science and Technology