Application of Predictive Analytics in Shipping and Maritime Logistics

Name: Jaimin Chainani

Candidate No: 298028

Supervisor: Prof. Jiabin Lou

Submission date: 12th September 2022

Submitted in part fulfilment of the requirements of Aston University for the degree of MSc. Business Analytics

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Prof Jiabin Lou, for her enthusiasm, patience, helpful insights, practical advice whenever I needed throughout the course of my dissertation. Her immense knowledge, profound experience and professional expertise has enabled me to complete the project successfully

I would also like to thank Dr. Leonidas Astanisakis, Dr. Viktor Pekar, Dr. Shahin Ashkiani and all of my module leaders for giving me strong theoretical and practical knowledge of Big data and Predictive Analytics concepts during my course that aided me in completing this business project.

Lastly, I would like to thank my family and friends for supporting and believing in me throughout the course. Their moral support provided me immense strength in order to overcome my academic challenges.

Executive Summary

Maritime and Shipping industry is the major route of trade across the world. According to United Nations Conference on Trade and Development (UNCTAD), Maritime Transport accounts for nearly 80% of international trade. In fact, UK ports facilitate approximately 95% of country's trade and play a significant role in the global supply chain. In 2019, there were over 95,000 ships in the waters and this number is rising significantly every year with stable increase in the total capacity as well. This increasing number is putting pressure over the industry to optimize the whole supply chain, increase the efficiency whilst reducing the risks.

The challenges can be overcome by embracing Digitalization. Thanks to the new technologies like Artificial Intelligence and Machine Learning, the shipping and Maritime industry can reconsider their approach by becoming more data driven. Considering port operations, even a slight improvement at strategic, tactical, and operational decision levels will lead to considerable cost decrease collectively. Machine Learning proves to be beneficial in various port operations like navigation and maritime safety, fleet management, Vessel turnaround time prediction, Border safety etc. In this thesis, we decided to dig deep and learn about trends and implementation of predictive tools in maritime industry.

Further after comprehensive research we tried to implement the techniques by studying the challenges of shipping in the Kattegat strait of Baltic region and we built a classification model using the AIS Dataset to predict the type of vessel flowing in that region. Several ML algorithms were considered to build and finally an accuracy of 98% was achieved by implementing a RF classifier.

Majority of Ships flowing in the waters of Kattegat region were found out to be cargo ships and tankers which are more hazardous to nature and can cause considerable damage to the marine ecosystem. Also, due to the geographical limitations of the region the Kattegat strait is more prone to accidents, and it is very difficult to navigate the large container vessels. Therefore, this model is aimed to recognise the patterns and improve the decision-making abilities in order to avoid maritime risks and optimize the supply chain

Table of Contents

List of figures/tables

Figures

Tables

List of Abbreviations/Acronyms

Acronyms	Abbreviations
AIS	Automatic Identification System
BAP	Berth Allocation Problem
DNN	Deep Neural Network
GDP	Gross Domestic Product
MAE	Mean Absolute Error
ML	Machine Learning
RMSE	Root Mean squared error
SVM	Support Vector Machines
VAT	Vessel Arrival Times

Chapter 1 Introduction

1.1 Maritime Industry in Global Context

Maritime transport is the backbone of international trade and the global economy. Around 80 per cent of global trade is carried by sea and is handled by ports worldwide (UNCTAD, 2018). Even in pandemic situations like covid-19, shipping has been crucial for ensuring supply lines globally and delivering necessary stores of food, fuel, and medical supplies from one location to another. The marine transportation network consists of specialised ships, the ports they go through, and the infrastructure for moving goods from production facilities, terminals, distribution hubs, and marketplaces. Ports and harbours are the main nodes in the maritime transportation network, which are connected through shipping routes (Han, 2018). Depending on cost, time, and infrastructure constraints, maritime transportation may be an alternative to roads and rail on various routes, such as certain coastwise or shortsea shipping or inside inland river systems. Other significant marine transportation activities include national defence, fishing and resource extraction, passenger transit and navigational services (James C, 2008, p. 6).

1.2 Digitalization in Maritime Industry

Like most other industries, shipping is undergoing a rapid transformation because of several technological advancements that aim to make operations more affordable, efficient, and green. (LAM, 2020). To increase productivity and operational efficiency at all decision levels, Investments in automation and digitalization of ports have expanded significantly (Zarzuelo, 2020). The advancements brought about by these investments have given rise to "Smart Ports" (Li Da Xu, 2018), a new phase of digital transformation for the port. Internet of Things (loT), Blockchain technology, Big Data, robotics, augmented reality, and Artificial Intelligence are just a few of the novel technologies that make up the most recent transformation wave of Industry 4.0. Industry 4.0, which relies on transformational cyber-physical systems, will become a critical element in both smart manufacturing and delivery and will transform the management of interconnected systems, thereby increasing competitiveness. (Digital, n.d.) Although the adoption of new technologies has accelerated recently, there are still significant

efforts to be made to achieve Industry 4.0's ultimate objective (Zarzuelo, 2020). By using such technologies at ports, a vast array of diverse data streams become accessible in ports, enabling the sector to tap into the power of data and optimise operations (Siyavash Filom, 2022).

Al is defined as "advancing the scientific understanding of the mechanisms underlying thought and intelligent behaviour and their embodiment in machines (Anon., 2018). The most significant branch of AI is machine learning (ML), which enables a system to automatically learn from data without explicit programming (Bhavsar, 2017). In numerous industries, ML approaches for data-driven decision-making showed promising results. Organizing the uses of ML and Al in port operations into categories could clarify earlier research and coordinate the field's future research plans (Siyavash Filom, 2022).

1.3 Problem Identification:

Context:

The European Shipping industry operates one of the largest, youngest and most innovative fleets in the world. The EU shipping sector adds a total of €147 billion to the EU's annual GDP3 with its wide fleet of container ships, tankers, passenger ships, bulk carriers, offshore service vessels, and many other specialised ships. Additionally, the fleet has one of the world's best safety records (ECSA, 2019-24). Accordingly, different needs of the European nations are met by different segments of this industry. Ferries that carry both people and commodities are an essential part of Europe's integrated transportation system, along with the short sea shipping sector they form a single market. Through a network of regularly scheduled services that link us to our trading partners, liners maintain Europe's trading capacity. Tankers, Dry bulk carriers, and LNG/LPG carriers ensure the safety and security of the supplies in energy, raw material and staple goods. By ensuring supply diversification, the EU shipping sector thus contributes significantly to preserving the EU's geopolitical independence (ECSA, 2019-24).

The Kattegat strait between Denmark and Sweden, which experiences high international vessel traffic, is no exception to the general trends (Lecq, 2021). The Baltic Sea is a hive of maritime activity, accounting for 15% of global cargo traffic, this strait connects the North Sea and the Baltic Sea and vessel traffic of nearly around 2000 ships can be found passing through its narrow straits at any given time (Davies, 2020). Cargo ships and tankers primarily use the established shipping routes. Not only are shipping routes through the Kattegat busier than those in the Baltic Sea, but almost no part of the Kattegat region is free of vessel traffic (HELCOM,

2018). High-speed passenger craft and ferries crossing the main traffic stream, a large fishing fleet, and smaller, seasonal craft are also relevant groups of vessels.

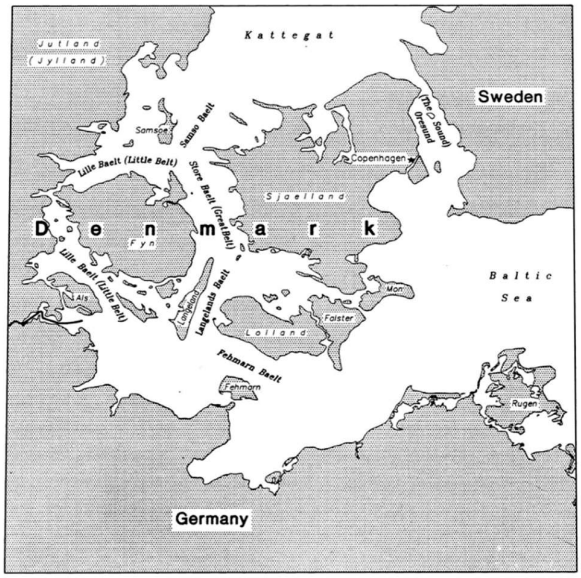

Fig: 1.1 Baltic region (International Institute for Law of the Sea Studies, 2021)

Difficult navigational conditions complicate matters for ships sailing through the Kattegat strait. Much of the water between Denmark and Sweden is less than 30 metres deep, limiting navigable space for deep-draft shipping (Pentti Kujala, 2020). As a result, there is a high number of groundings (HELCOM, 2018). Furthermore, Kattegat's unique environmental configuration has an impact on navigational safety. High vessel traffic densities and difficult navigational conditions make the Kattegat susceptible to ship collisions. The midwater phase of a ship's journey within internal waters or on territorial sea is regarded as the most dangerous. In the last decade, there has been no significant decrease in the number of maritime accident events.

Hence, this research study is aimed towards getting an overview of the newest technological trends in the maritime industry. Further, emphasis on the practical implication of machine learning tools would strengthen the research. In the latter part, we would build a predictive model to predict the type of ship in the Kattegat region, using an open-source AIS (Automatic Identification System) dataset derived from Denmark Maritime Authority. Larger ships with deep draughts pose a significant issue, particularly for routes that pass through the shallow sections of the Baltic Sea and for port development because channels must be deeper and wider. therefore, Our research outcome is focused to add value to ML applications in port operations

like demand forecast, navigational safety, seaside/portside operations, vessel turnaround time prediction etc.

1.4 Research Objective

The objective of this master's Dissertation is fourfold:

- To gather historical data and learn how knowledge can be gained from the data already available
- To identify techniques and build a model, useful for making predictions based on historical AIS data
- To evaluate the model's reliability through testing and comparison.
- To suggest a prototype that can be incorporated by already existing stakeholders for optimization and safety of shipping industry

Therefore, we summarise our chapter 1, In the initial part of the chapter we studied about the importance of maritime in global context and the current trends in Digitalization. Thereby, we put forward and set up the premise for the focus of our study and presented our research goals. In the next chapter we will study about concepts and theories relevant to our research.

Chapter 2 Literature review

We begin this chapter 2, by giving an introduction of the concepts of data mining and machine learning techniques relevant to conduct our research. Further, we provide an overview of digitalization in maritime industry and merits of machine learning in shipping and maritime supply chain. After conducting proper research, we provided an overview of the application of ML techniques specifically for improvements in different areas of port operations. Thereafter, we directed our focus of the study towards the advancements in seaside operations of maritime supply chains. To validate our research, in sections 2.6 and 2.7, we introduce the maritime challenges in the Baltic region and navigational difficulties in the waters of the Kattegat strait region near Denmark and conclude this chapter.

2.1 Data Mining and Machine Learning

2.1.1 Origins

Data mining originally is a process of sorting through large data frames and digging out patterns and relationships that can aid in the solution of complex business tasks and improve decision-making abilities through thorough data analysis. Data mining is a term which is commonly accompanied by the concepts of Statistics, Machine Learning and Artificial Intelligence (Pang-Ning Tan, 2006) (Johansson, 2007).

According to (Kantardzic, 2011), data mining has its roots in statistics. On the other hand, Statistics has its origins in mathematics, thus it has a desire to test anything theoretically before putting it into practice. The machine learning methods share their roots in computer science, which has a more practical perspective as compared to mathematics. Anything is tested and evaluated to see how it performs without necessarily having a formal demonstration of effectiveness. Machine Learning includes the word learning which implies a process, process is instead an algorithm hence, machine learning emphasizes algorithms whereas statistics focuses on models. (Pang-Ning Tan, 2006) explains in their research that data mining is the synergy of several disciplines and draws on statistics, like sampling, hypothesis testing,

estimation etc. and on the other hand learning theories from artificial intelligence, and Machine learning ((Nina), 2019)

2.1.2 Purpose

Further (Pang-Ning Tan, 2006) (Johansson, 2007), describes the goal of data mining, to either describe certain patterns or predict specific values. The descriptive goal is to obtain an understanding of the analysed data by identifying patterns and relationships and the predictive goal includes building a model that may be used for classification, estimation, or other similar tasks. Therefore, a combination of both the fields can be helpful to solve complex data mining tasks and challenges, as it requires to first search for patterns in the data and then use these patterns to build a predictive model ((Nina), 2019)

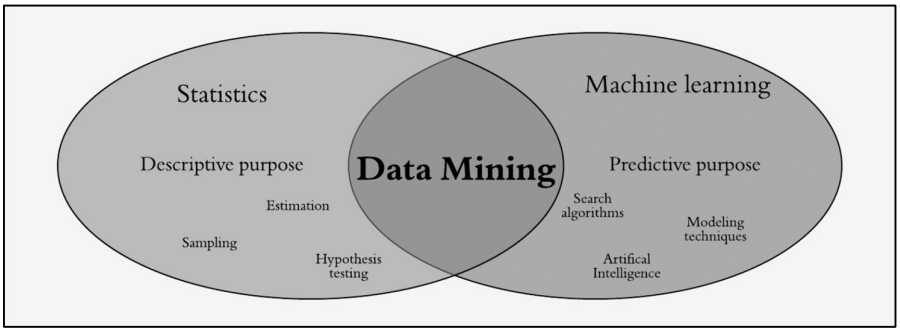

Fig: 2.1 Data mining perspective ((Nina), 2019)

2.1.3 Predictive Analytics

Predictive data mining, according to (Freitas, 2002), it refers to forecasting the value of a target variable in the future based on previously observed data. In simple words, predictive analytics refers to combination of statistics and modelling techniques to make predictions about future outcomes and improvements in performance. If a target value comprises of numerical values, then the problem is known as regression. Similarly, if the target variable includes discrete class labels, then it is known as classification problem. Both can be used to generate a model that minimises the difference between the predicted and true values.

Classification and regression are both concerned with the problem of training a target function that correlates each variable set to one of the predetermined labels in order to categorise unseen instances ((Nina), 2019). The nature of our research study involves a classification problem. In Chapter 3, we further understand the theories and concepts and build a predictive model and in

chapter 4, we evaluate the model and understand the specifics which further are required to improve the model's accuracy.

2.2 Maritime Connectivity:

Machine learning techniques based on data are built on the assumption that a large amount of data is available in abundance (Lutz Kretschmann, 2019). Historically, the marine industry faced a key challenge of lack of adequate connectivity or the inability to transfer greater data volumes at all. Therefore, limited data accessibility from shore, narrow bandwidth, and in some cases, ignorance of the significance of the details and information evident in the data placed restrictions on innovative solutions. Due to the recent major advancements in maritime connectivity, it is predicted that this trend is rapidly changing and is anticipated that this trend will continue in the future. Over the ocean, data exchange is obviously only possible via satellite communications, which makes data exchange even more expensive. Higher data transmission speeds and lower communication costs are anticipated, providing opportunities for innovative applications that require the frequent exchange of large amounts of data between ship and shore (Wingrove, 2020).

Another accomplishment in the digitalization of marine operations is the commitment to equip vendor ships with Automatic Identification System transmitters (AIS). Since the introduction of AIS, the computerised exchange of position, speed, course, and other information via AIS from one boat to another or from boat to shore station has significantly increased both the productivity and security of oceanic traffic (Lutz Kretschmann, 2019). At the same time, the availability of AIS data over many years has generated a unique database that reflects the geographical movement patterns of the whole merchant fleet and the navigational behaviour of particular ships. (International Telecommunication Union, 2014)

Overall, improvements in the maritime sector are paving the way for seamless connectivity between the water and the land as well as for the real-time sharing of massive and intricate maritime data sets. All of this makes it possible to apply machine learning technologies in maritime logistics more frequently (Lutz Kretschmann, 2019).

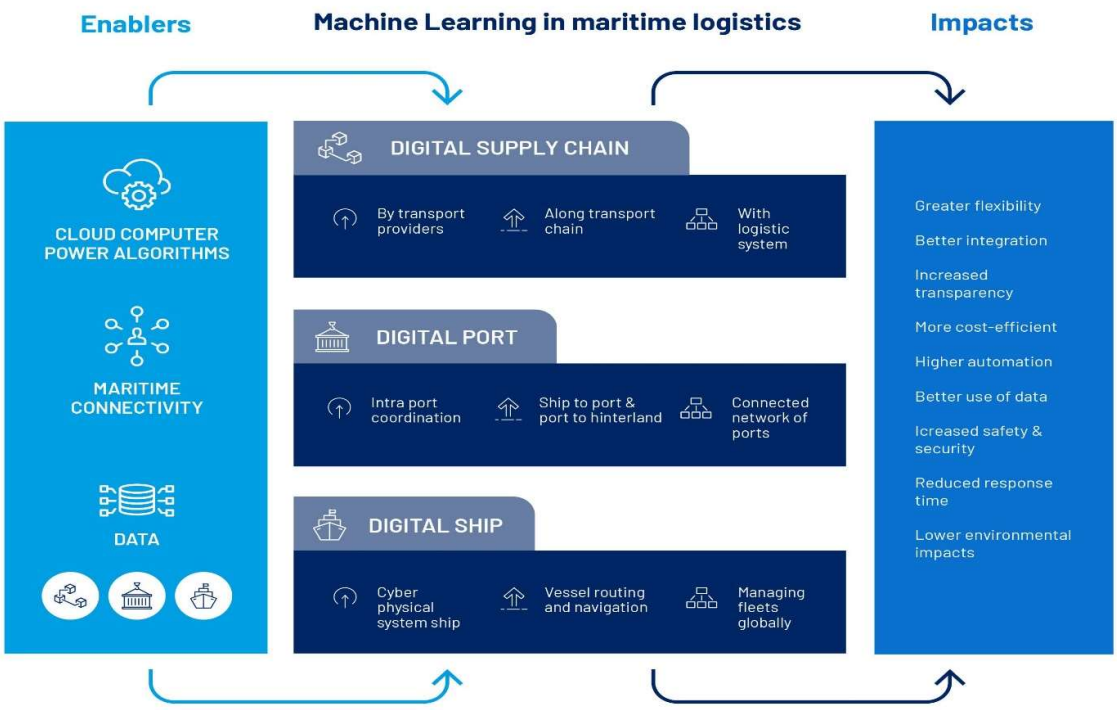

Fig 2.2 Impact of Machine learning in maritime logistics (Lutz Kretschmann, 2019)

2.3 Digital Ports and Supply Chains

Digital Supply Chains can help maritime logistics reach a higher degree of operational efficiency by making use of technology like the Internet of Things and advanced analytics. Real-time observation of the flow of commodities, accurate projections of their movement, and automated and data-driven decision-making are a few of the important functionalities of digital supply chains (Lutz Kretschmann, 2019). Shipping companies, freight forwarders, and transportation firms would all profit by providing more dynamic, cost-effective, and well-coordinated transport services. Here, technological advancements and global connectivity allow for smooth information sharing throughout a shipper's international supply chain. By taking advantage of the massive amounts of data that cargo monitoring devices like AIS offer and extracting meaningful information from it, machine learning can be extremely helpful in achieving enhanced visibility of individual shipments. Additionally, data-driven decision-making, made possible by digital supply chains, can far more effectively address the inherent complexity and dynamics of logistics processes than existing approaches. At the same time, new technologies make data sharing tamper-proof and traceable, assisting in ensuring security in digital supply chains (UNCTAD, 2019).

The physical handling of commodities, loading devices, means of transportation, and information technology to optimise port calls are all examples of port-centred developments. Extending the planning horizon of port stakeholders through intra- and inter-port collaboration, ship-to-port collaboration, and port-to-hinterland collaboration is particularly significant. As a result, vessel turnaround times are shortened, and resources are used more effectively (Sea Traffic Managmenet , 2019). Additional advantages are promised by synchronising this kind of international shipping across numerous smart ports. It is possible to better coordinate handling and transport capacity through information exchange and data-based predictions, which creates prospects for cost savings and efficiency gains along digital supply chains (Lutz Kretschmann, 2019).

As more data becomes available, developments in other industries have proven that digitization is not only fundamentally altering corporate processes but also producing completely new business models. Similarly, in the maritime sector, there is no exception; the logic of the current sector is already being disrupted by new data-based services and solutions. These new digital functions and services offered by ship managers, the shipbuilding and marine equipment sector, classification societies, and other maritime service providers frequently heavily rely on advanced analytics and machine learning (Lutz Kretschmann, 2019).

2.4 Opportunities and Limitations: -

According to the current industry trends, it is important for a maritime company to leverage the value of data with artificial intelligence and machine learning tools in order to gain and maintain a competitive advantage in all maritime industry segments, from shipbuilding and the maritime supply industry to offshore wind and marine engineering. The key opportunities include:

- Reduce costs by optimising operations and making data-driven decisions.
- Enhance quality control with digital monitoring solutions.
- Increase safety by predicting incidents along with autonomous/remotely controlled operations
- Extract and deduce more information from the historic records
- Recognise decision-relevant information through large data sets, thus, reducing the time required for document handling and processing.
- create new business models and products by automating processes with intelligent assistants

Besides opportunities for new technology, there are several barriers which limit the functionality of the technology. Some of the challenges faced by companies presently are as follows:

- Inadequate data quality can result in a large amount of data cleaning and pre-processing work.
- Expectations with computer systems can reach unsustainable heights, thereby putting human effort and intelligence in jeopardy
- Large computational power needed for handling and processing large datasets

2.5 Application of ML techniques in port operations

Artificial intelligence is a critical technology for Smart Ports and the driving force behind port automation. Smart Ports could not exist without Artificial Intelligence, AI has already had an impact on global logistics companies and will continue to have an impact on the maritime and shipping industry's development. Machine learning, a subset of AI in which machines are programmed to replicate and imitate human decision-making processes, assists a business in creating a digital simulation of what might happen in the real world (Sinay, 2021). In this section, we will review and discuss the application of ML techniques in port operations. For a more systematic review, the applications are classified into the following categories:

a) **Demand Prediction**- According to (Siyavash Filom, 2022), the use of ML in a port's overall demand forecasting differs from the other areas of the marine industry. Demand forecasting applications concentrate on predicting port throughput, which is not considered an "operation" but has a significant impact on port operations. For port operators and stakeholders, accurate demand forecasting can be a crucial asset that offers both immediate and long-term benefits. In the near future, virtually all port actors, including decision-makers, terminal operators, hinterland service providers, and others, will be able to plan their tasks more precisely and effectively, thereby resulting in smoother port operations and increased efficiency. Reliable demand forecasting, on the other hand, enables the port stakeholders to make effective long-term strategic decisions like port expansions. For port expansions, it is necessary to strategise properly as it involves a lot of complex decision-making and one mistake can result in some

irreversible financial decisions. Thus, it should be based on accurate and concrete port demand forecasts.

Container throughput is considered one of the most significant indicators to measure the development of port economy. It is an inherently complex and dynamic process because of numerous socioeconomic factors influencing the outcomes such as Gross Domestic Product (GDP), location, seasonality patterns, fuel price, population, policies, and political tensions (Wen-Yi-Peng, 2009) (Shankar, 2020). Most of the demand forecast research studies are based on historical port throughput time series data. Historically, time series problems have been solved by employing statistical models that are primarily based on linear assumptions (An-sing Chen, 2004). Recently, in lieu of the traditional linear statistical models, ML algorithms have been successfully established (Siyavash Filom, 2022).

b) **Landside Operations-** Operations on landside begin when cargo is loaded or unloaded for a ship at berth and terminate when the ship exits the gate. Any port's main operations are transportation, storage, and loading and unloading. Sometimes, Uncertainties in demand forecast and short time intervals between arrival ships results in demand more than terminal capacity. The emerging ships with giant dimensions also impact the port's terminal operations (claudia caballini, 2020). The term "terminal operations" refers to a variety of activities carried out in ports, from the quay—the point where land and water meet—through the final delivery of cargo. This operational flaw in the terminal creates a bottleneck for overall port operations, potentially affecting the ports' competitiveness and causing financial losses (Siyavash Filom, 2022). There are several areas of application of ML in landslide operations such as:

    a. **Stowage planning-** In easy words, stowage planning is the act of allocating space on the vessel for cargo that must be loaded from a specific port or ports to be discharged at a certain port or ports without those containers having to be handled again at any of the other ports along the route. Loading containers from a port involves an important series of decisions that port operators must make, which affects ship turnaround time. If the port operators fail to adhere to the time frame for operations, the ship charterer is liable to pay demurrage to the ship owner, lowering the port credit and raising the cost of logistics. As a result, optimising the "stowage plan" is critical for port operators.

b. **Container Dwell Time**- Another key port efficiency indicator is "container dwell time," which shows how long a container is stored at the port. Higher dwell indicates inefficiency, which lowers overall port productivity (Nadereh Moini, 2012). (Ioanna Kourounioti, 2016) tried to predict the dwell time o fimport containers in the middle east. They used deep learning methods to predict dwell days based on terminal data which resulted with prediction accuracy level of 65%. Predictive and prescriptive analytics are just two of the many ML applications used in landside operations, which are diverse due to the operations' wide range.

c) **Seaside operations-** It is essential to emphasise that there are numerous applications of ML methods in ship and sea operations, many of which are stimulated by rich AIS data (Dong Yang, 2019). The majority of them can be classified as autonomous shipping (Chen, 2016) (Trudi Hogg, 2016) (Steven C. Mallam, 2020), vessel route planning and fleet management (Marielle Christiansen, 2013), collision avoidance (Yamin Huang, 2020) (Kadir Cicekc, 2019), environmental shipping evaluation (Abebe, 2020), ship traffic patterns, ship fuel consumption (Ran Yan, 2021), and anomaly detection systems (Riveiro, 2018). This part comprises and discusses articles related to "Port Seaside Operations." Berth operations (managing berth and ship-to-shore crane resources), approach channel and basin operations, and vessel operations within the port area are all part of port seaside operations. Berth Allocation Problem (BAP) and Vessel Arrival Times (VAT) prediction are two traditional problems in port seaside operations (Siyavash Filom, 2022).

a. **Berth Operations**- The BAP problem seeks to strike a balance between inbound vessel arrivals at terminals and a limited number of quays in order to reduce the time ships spend at the port. Combinatorial optimization with many physical, technical, and operational limitations is frequently used in BAP (Ching-Jung Ting, 2014).

b. **Basin and approach channel operations**- The port's wet infrastructure, which includes the port approach channel and harbour area, is one of the most important assets of any port, with capacity representing the overall port capacity in many cases. Therefore, acquiring an appropriate estimation of the approach

channel density and prediction of the vessel movements in the harbour area could lead to more robust vessel tactical planning (Siyavash Filom, 2022). One of the most common bottlenecks at ports is channel capacity, thus port managers must estimate channel capacity realistically. At the Port of Tianjin, (Cong Liu, 2020) tried to determine the navigational capacity of the Dagusha channel. The K-means clustering method is used to cluster ship traffic based on ship type using one month of AIS data. The outcome determined the maximum channel capacity, which helps with vessels' arrival and departure planning.

c. **Vessel turnaround time**- Vessel turnaround time is an important performance indicator in port operations. Berthing time, waiting time, and service (loading/unloading) time are all part of this parameter. This indicator has a significant impact on the overall effectiveness and capacity of port operations (René Taudal Poulsen, 2020). It must be noted that longer turnaround times cause several issues, including ship demurrage, increased vessel emission levels in the port area and decreased port competitive strength.

d) **Safety**: Naturally, safety procedures and measures are becoming increasingly important in port operations. Ports are complex organisations with multiple socioeconomic and environmental factors interconnected. Similarly, any safety issue could cause serious economic and environmental problems, financial and human loss, impact the port's competitiveness, and potentially puts the port out of service.

a. **Navigational Safety-** Listed below are a few safety risks associated with vessels sailing within the port area (basin + approach channel). Even a minor incident could pose a significant challenge to the port and freight lines. A small instance can cause a port entrance to be blocked or port facilities to be damaged, resulting in cascading effects on supply chains and a loss of competitiveness. The recent suez canal incident is the perfect exemplification which resulted in huge losses and involved serious repercussions to the stake holders.

Using ML methods to prevent such incidents could provide useful feedback and predictions aimed at preventing or reducing risks. (Kadir Cicekc, 2019) investigated navigational collision risk in port approaches and basins that are

critical to port authorities. By investigating 140 pilot approach manoeuvres in a ship handling simulator, a notable dataset was gathered. Then, 20 expert pilots were interviewed to extract a set of collision prevention fuzzy rules. The subsequent datasets were fed into a random forest, which produced superior results.

b. **Port State Control-** The port state inspection, which is carried out by the port authority to protect the safety of ships, crew, and the maritime environment, is one of the most crucial safety procedures in ports. The port state inspection, which is carried out by the port authority to protect the safety of ships, crew, and the maritime environment, is one of the most crucial safety procedures in ports. If a ship cannot fulfil the required criteria, it gets grounded. Age, type, deadweight, and registered flag state of the ship are some of the factors that port managers must consider when deciding which ship to choose for PSC inspection (Christiaan Heij, 2019) (Yi Xiao, 2020).

2.6 Shipping in Baltic Sea:

The Baltic Sea region is one of the most important regions in the maritime industry, with about 2000 ships passing at a time through the straits of the region (Davies, 2020). Thus, making it one of the heavily trafficked seas in the world. The Baltic's narrow straits and shallow waterways, most of which are covered by ice for extended periods during the winter, make navigation difficult and increase the risk of maritime accidents (HELCOM, n.d.). In the report (Interreg Baltic sea region, 2016) the authors provide insights about the growing container and cargo vessels in the Baltic region. The reports suggest that general cargo ships account for more than half of the total ships. Approximately 20% of the ships were tankers transporting more than 200 million tonnes of oil, while there are only 11% passenger ships carrying approximately 50 million passengers. Despite increased trade volumes, the number of ships transiting the Baltic Sea has declined, indicating a trend toward larger vessel sizes. These gigantic vessels can greatly be impacted by the difficult weather conditions and the winds flowing in that region. To monitor and maintain the safety at sea, the organization HELCOM annually compiles shipping related accidents in the Baltic Sea

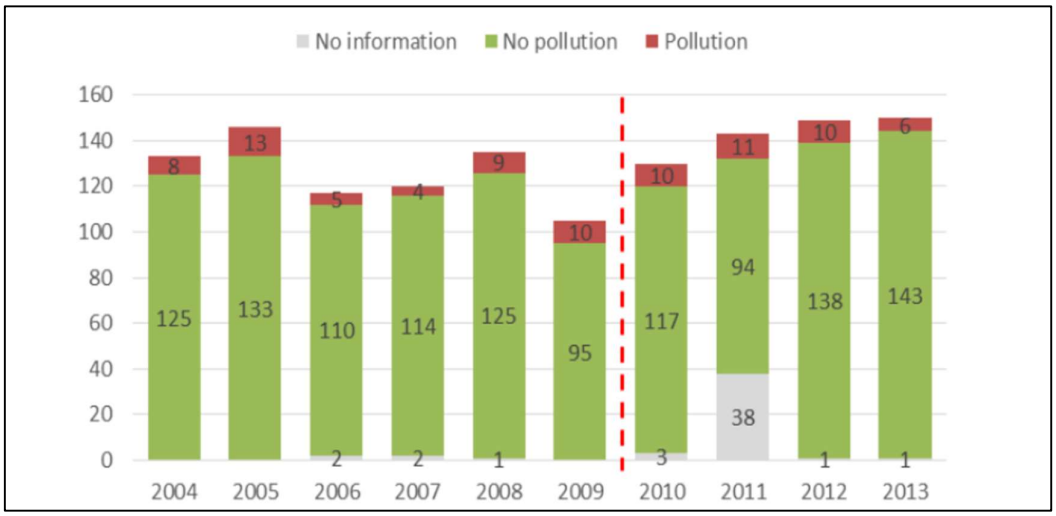

Fig 2.3: Number of reported ship accidents from 2004 to 2013 (Ernest Czermanski, 2018)

In the year 2013, 150 of such incidents were recorded, majority of them were cargo vessels. It was also reported that 28% of the accidents were due to human errors followed by technical failures at 19 % (Interreg Baltic sea region, 2016). Thus, the figures indicate and alert the ports in the region to plan strategically and invest their funds in strengthening their technology and infrastructure in order to provide access to larger ships smoothen the port operations.

2.7 Maritime in Kattegat Strait

The Kattegat region is no exception to these Baltic trends. The Kattegat strait between Denmark and Sweden feature a lot of ship traffic. This strait connects the North Sea and the Baltic Sea, and over 75,000 ships passed through it in 2019 (A Grimvall).The Baltic Sea consists of the Little Belt, Great Belt and the Sound, the shortest area between the Baltic Sea and the Kattegat and the North Sea (International Institute for Law of the Sea Studies, 2021). Its narrowest point is 2.2 miles wide. However, it lacks the depth necessary for deep-draught vessels. Only deep-water channel available spans the Great Belt, which is 10 miles wide (International Institute for Law of the Sea Studies, 2021). These established channels are typically used by cargo and passenger ships. Not only these shipping routes witness heavier vessel traffic than Baltic Sea, but there is also hardly any area free of ships (HELCOM, 2018). Ships transiting the Kattegat face an additional layer of complexity due to challenging navigating circumstances. The depth of the sea between Sweden and Denmark is generally less than 30 metres, which restricts the amount of navigable space for deep-draught ships (Pentti Kujala, 2020).

Fig 2.4 The three belts of Kattegat region (International Institute for Law of the Sea Studies, 2021)

The Kattegat's unique environmental Factors has an impact on navigational safety. In the region, 474 maritime casualties and incidents were reported between 2011 and 2018. Between 2000 and 2017, the IMO (2017a) recorded 20 collisions and groundings. The waterways of Baltic Sea are getting even busier with the steep rise in transportation of oil. Even the region is becoming increasingly popular as a cruise destination, which has increased the number of passenger ships exploring new routes recently. This has sparked worries about the increasing likelihood of maritime accidents (Davies, 2020)

Here, we conclude the literature review relevant to this study. Initially we introduced the concepts of Data Mining and impacts of Machine Learning on Shipping industry and also, we took in account the opportunities and limitations of newer technologies. Also, we took a bird's eye view of applications of Machine Learning in Port operations. Thereafter, we identified the gaps and difficulties of maritime in Baltic region and the need for a better predictive model to resolve the conflicts occurring in the industry effectively. In the next chapter, we touch upon the methodologies used to conduct our research.

Chapter 3 Methodology

3.1 Methodology overview

In this part, we describe our approach for building a model to predict a ship type sailing in the waters nearby the Kattegat Strait region in Denmark. This model can be used by various stakeholders involved in the supply chains management of maritime in the Baltic region to ensure the smooth flow of ships, thereby reducing risks and optimizing various seaside operations. The analysis is based on the data mining techniques and hence, The Cross-Industry Standard Process for Data Mining (CRISP-DM) method, which is considered as the most widely used in practise, was chosen to conduct the research (Pete Chapman (NCR), 2000). It offers a structured method for organising a data mining project, with six phases that are roughly depicted in figure below. While the input for each phase dictates the output of the one after it, the order of the phases is flexible, so switching back and forth between them is always necessary ((Nina), 2019).

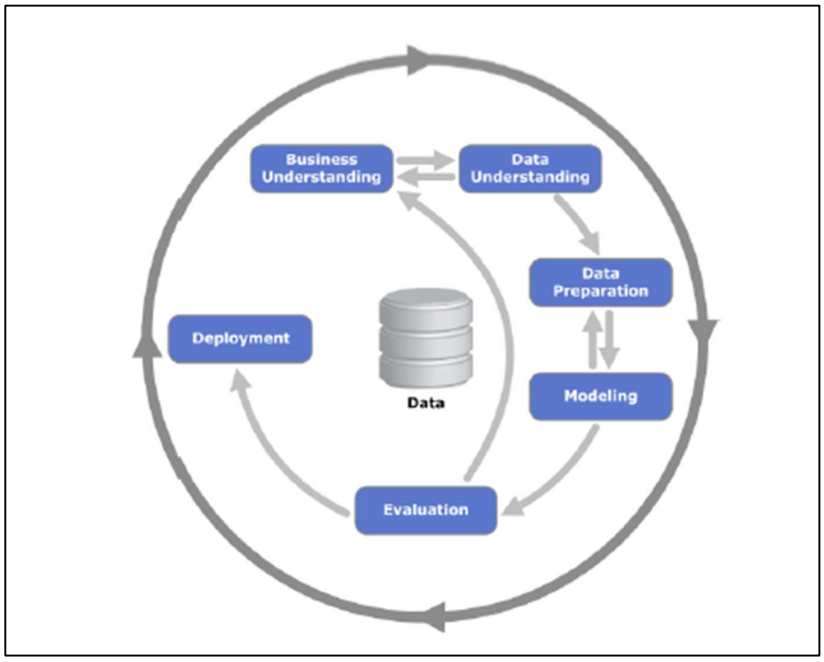

Fig 3.1 The CRISP DM Cycle (Pete Chapman (NCR), 2000)

For acquiring the data required for analysis, we use satellite data from the automatic identification system to establish a solid foundation for the prediction of ships. An automated identification system (AIS)for ships is used primarily to increase navigational safety, ship-to-ship communication, and ship reporting. Ports and other maritime agencies utilise AIS, which

is non-public information, to track vessels. Several Satellite data service providers like Marine Traffic provide access to AIS data through paid services.

3.2 Overview of ML methods

Machine learning pulls information from data generates accurate predictions and judgements on what has been learnt without requiring any prior knowledge of data or context. ML does not rely on existing rules or equations as a model, in contrast to the previous concept of intelligent systems, such as data mining or expert systems, which were based on pre-determined rules to interpret the data (Siyavash Filom, 2022).There are several fundamental types of learning techniques, including:

(1) supervised learning, in which the learning process is guided by previous data points.

(2) unsupervised learning, in which only unlabelled data is used.

(3) semi-supervised learning, in which both labelled and unlabelled data are used.

(4) reinforcement learning, in which the learning process is governed by a sequence of feedback/reward cycles (Bhavsar, 2017).

3.2.1 Supervised Learning

A function (or algorithm) is trained using supervised learning techniques to compute output variables based on given data that contains both input and output variables i.e The goal of supervised learning is to use a labelled dataset as a training set to learn how to predict new, unforeseen data with accuracy (testing set). Depending on whether the objective is to predict a class label or a numeric value, supervised learning techniques can be divided into two categories: classification and regression. There are several supervised learning techniques exist, the majority of which may be applied to both classification and regression problems. The most popular SI techniques include K-Nearest Neighbours (KNN), linear models, Naive Bayes (NB), ANN, Support Vector Machines (SVM), and tree-based algorithms.

3.2.2 Unsupervised Learning

Unsupervised learning aims to understand data by extracting features, co-occurrences, and underlying patterns without the use of labelled data instead of inferring models for input-output pairs. Data transformations, association, autoencoders, clustering, anomaly detection, and association are just a few of the uses for unsupervised learning. The major techniques used in unsupervised learning are principal components analysis and cluster analysis. Principal

Component Analysis is a popular technique for dimensionality reduction and exploratory data analysis It can help identify the correlation between features and works by finding the maximum variance in high dimension data and projecting this to fewer dimensions. When having many correlated variables, a PCA can explain this with fewer dimensions (Pena, 2020). However, clustering algorithms attempt to divide the data into groups where the data points are like one another but distinct from those in other groups.

3.2.3 Reinforced learning

The final type of learning problem is reinforcement learning. Reinforcement Learning allows an algorithm to learn through continuous trial and error. A reinforcement learning algorithm may include several elements, such as a reward signal, a policy value function, and, in some cases, an environment model. The policy defines the model's actions, the reward signal defines the learning problem's goal, and the value function represents the total reward over time (Pena, 2020). Learning the correct sequence of motions to complete a task in robotics is an example of applied reinforcement learning. Furthermore, the approach is not only prominent in the development of AI systems capable of outperforming humans in games such as chess, but also increasingly in technical or business-related settings (Lutz Kretschmann, 2019)

3.3 Data collection

3.3.1 Automatic Identification System:

An information-delivery system for ships called AIS was adopted by the International Maritime Organization (IMO) with a focus on maritime safety. Despite previous regulations, the use of AIS became mandatory for all vessels in 2004 through the Safety of Life at Sea (SOLAS) agreement. When a vessel is within the range of a transmission, AIS data can be used as a medium of data to spatially describe vessel movements. According to the Figure, a ship's onboard transceiver, ground/satellite station receivers, and vessel traffic services terminals make up the Automatic Identification System (AIS), a global, autonomous tracking system. The transceivers automatically and routinely broadcast AIS messages. It is possible to visualise the information obtained and utilise it to guide navigational decisions. The broadcasted AIS Messages typically contains two types of information.

- **Static Information-** contains ship name, ship MMSI ID, message ID, ship type, ship size, current time).
- **Dynamic Information**- comprises of ship location, speed, course, heading, rate of turn, destination and estimated arrival time.

Two kinds of shipborne equipment are identified by the IMO (2015), with each class having unique technical requirements and intended uses. All vessels above 300 gross tonnes when on international trips, all cargo ships exceeding 500 gross tonnes while not on international voyages, all sized passenger ships, and fishing vessels longer than 15 metres must be equipped with Class A equipment. This kind of technology is more expensive, highly developed, and intended to use commercially.

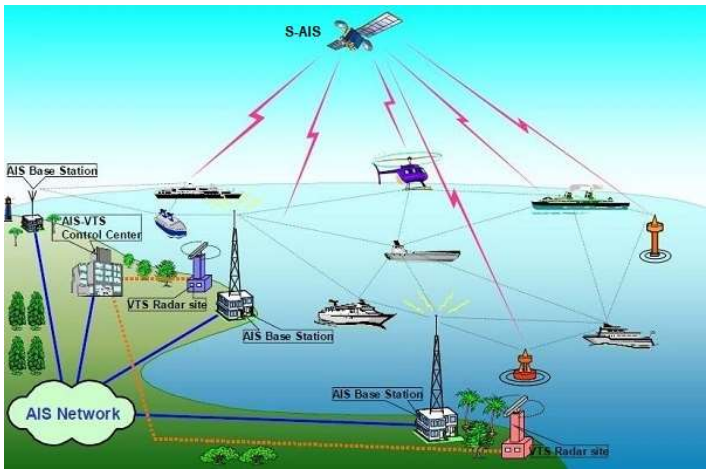

Fig 3.2 Automatic Identification system (NATO shipping centre, 2021)

Class A units must report every 180 seconds at the most when they are at anchor or moored and moving no faster than 3 knots, and every 2 seconds when they are travelling more quickly than 14 mph and changing direction. Vessels that are exempt from class A restrictions are equipped with class B transceivers. These transceivers are often less sophisticated, less expensive, and designed for smaller, seasonal boats and lighter commercial vessels. Class B devices typically have a reduced reporting frequency and a transmission interval of between 180 and 30 seconds. The messages can also be received by a satellite, which has a range of more than 400 nautical miles, or by an onshore station with a usual range of roughly 60 nautical miles. These AIS data can be transferred over long distances and stored in large quantities, and they have a high value for maritime data mining and intelligent navigation.

3.3.2 Application of AIS Data in maritime Industry

1. AIS data can be utilised to verify or contextualise other data. For instance, AIS could act as actual truth to verify or complete data from radars, more precise synthetic aperture radars (SAR), tracking radars, or satellite photos. Additionally, AIS data could be used to supplement other data sources, such as those that deal with geography or underwater noise. Additionally, AIS data might make weather-related information easier to understand. Consider the effect of weather factors on vessel behaviour, such as wind, visibility, currents, waves, ice, hurricanes, and/or season. As a result, this might have a significant positive impact on both the management of vessels and the environment.
2. AIS was first developed for vessel traffic safety and to aid in preventing vessel disasters, similar to the way GPS is used for the same purpose. Numerous research refer to this original risk- and safety-related goal. In the event of a collision, AIS data may potentially be used as legal evidence. Additionally, AIS data may be useful for preventing not only vessel crashes but also other kinds of collisions (such as vessel into land). Additionally, AIS data can be used to learn more about how different ship captains steer their vessels. The detection of piracy can also be aided by AIS data, In some of these situations, AIS data might assist in locating illegal activity by identifying "anomalies" of tracks.
3. Data from AIS systems can be used to control the environment. AIS data may generally be used to assess compliance with environmental regulations. For instance, it is possible to keep an eye on whether or not certain locations are prohibited or protected. By measuring, estimating, and forecasting the emissions of ships using AIS data, the quality of the air may also be preserved. In addition, the oil spill can be located. Additionally, by identifying regions where animal territory and vessel routes cross, AIS data can aid in the protection of wildlife.
4. AIS data can be useful as a strategic planning tool, when combined with Geographic Information system, data mining and other databases. In general, the utilisation of AIS data allows for the mapping of ship routes, identification of travel patterns, effective traffic management, retrieval of lost tracks, strengthening of prediction accuracy, and estimation of the basic ship traffic diagram through mining of ship speed-density relationships.

However, massive data modelling of historical AIS for intelligent navigation remains a difficult task. (Enmei Tu, 2020) describes the challenges with historical AIS trajectories by showing a small part of AIS database of vessels near the west coast of USA. Each dot in the figure represents an AIS message, and each line or curve represents an AIS trajectory segment of a vessel. From the figure, it can be inferred that the AIS has the following issues:

- The trajectories vary significantly in length, shape, position, and orientation
- A learning algorithm may be misled by the "abnormal" vessel motion patterns that regularly appear in AIS trajectories (such as wavering, U-turns, and self-intersection), which lowers the algorithm's capacity for generalisation.
- The message frequency in AIS trajectories varies, meaning that certain trajectories may have dense message sequences while others may have relatively sparse message sequences, sometimes with missing data and incorrect values.

Moreover, each AIS trajectory also includes both static and dynamic data. These problems and the abundance of big data present a significant challenge to current methods.

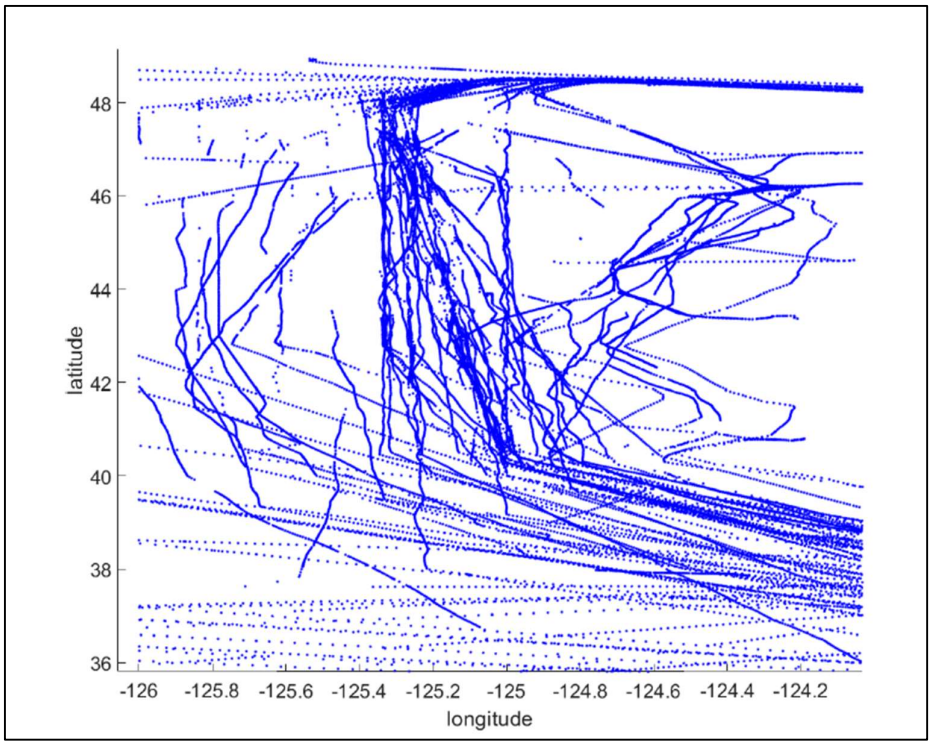

Fig 3.3 Example of AIS Trajectories (Enmei Tu, 2020)

3.4 Source of Dataset

There are certain challenges associated with sourcing the AIS dataset for research purposes. As mentioned earlier in the research, the Automatic Identification dataset is a highly complex and dynamic set of information which is updated every second. Hence, a certain kind of high-quality infrastructure of computing power is required in order to use this information. Big Maritime Companies and stakeholders involved in the supply chains of the shipping industry possess such resources and infrastructure to imply and use this dataset for their logistics management. Naval Authorities and the Government use this data for guarding their ports and resources from maritime risk.

As this data is not available for public use, there are rare chances of finding an open to use, reliable AIS dataset for research purposes. As previously addressed, several satellite data service providers provide access to AIS data through paid services. The current Automatic Identification Data systems which are available for public use are struggling with reliability of the data. In (Ties Emmens, 2021), the authors validate by stating that the static, dynamic, and voyage-related information such as speed over ground, timestamps, position, etc. are communicated incorrectly and hence contain a lot of noise. Human errors are also one of the reasons for data redundancy or missing information.

According to (Abbas Harati-Mokhtari, 2007), 80 percent of the data is inaccurate Similarly, there were subsequent difficulties in finding an appropriate data for performing the analysis to validate this research. Hence, due to the limitations and the challenges encountered in data sourcing, an open-source online platform named Kaggle was finally considered for the purpose of this study. Kaggle is a place which enables users to publish their work and search for open to use datasets for machine learning projects.

To make it more relevant to current statistics and adhere to the time boundaries, an open to use processed data sourced from Kaggle bur originally a part of a dataset published by Denmark Maritime Authority (DMA) was considered for the analysis. The dataset comprises of two months of vessel traffic data i.e from January March 2022 sailing in the region of Kattegat Strait located in the Danish Straits Islands, North of Baltic Sea in Europe. The data was downloaded in .csv format which was imported in python to perform the analytics

3.5 Preparation of Data

Software Used

The data analysis for the research study was carried out with the help of python. Traditionally the machine learning tasks by manually coding all algorithms, mathematical and statistical formulas which increased the complexities. But recently with the newer developments, this complexity has decreased significantly. Today, Python is one of the most popular programming languages for Machine Learning problems. Due to python's simplicity and consistency and large library of tools, it has replaced many existing programming languages in the industry. Some of the python libraries used in this machine learning project are:

- **NumPy**- NumPy is a prevalent Python library for analysing large multi-dimensional arrays and matrices using a large collection of high-level mathematical functions. It is extremely useful for fundamental scientific computations in Machine Learning.
- **SciPy-** SciPy includes several modules for optimization, linear algebra, integration, and statistics, it is a widely used library amongst machine learning enthusiasts
- **Scikit-learn**- Scikit-learn is developed using the two fundamental python libraries NumPy and SciPy. The majority of supervised and unsupervised learning algorithms are supported by Scikit-learn
- **Pandas-** Pandas were built specifically for data pre-processing and data extraction. It offers a high level of data structures and a large range of tools for data analysis
- **Matplotlib**- matplotlib and seaborn are python libraries used for data visualization. It provides various kinds of graphs and plots viz. histogram, error charts, bar charts etc.

The above collection of libraries were helpful in conducting our project. In the next chapter, we conducted an exploratory research over our dataset. We understand the patterns and key relationship of between significant variables and we conclude the chapter by explaining the data cleaning and feature selection process and prepare our model ready for model development.

Chapter 4 Data Analysis

4.1 Description of Data

It is important to understand the data before performing the operations. By describing a sample of the data set, we get to know certain information about the variables and their types. Below is a snapshot of the first 10 values of the original data set collected for the analysis. There are in total 10 variables and 358351 rows of information.Without performing any further operation, we can infer that there are certain columns with missing values.

	A	B	C	D	E	F	G	H	I	J
		mmsi	navigationalstatus	sog	cog	heading	shiptype	width	length	draught
0		219019621	Unknown value	0	86	86	Fishing	4	9	
1		265628170	Unknown value	0	334.5		Port tender	8	27	
2		219005719	Unknown value	0	208.7		Fishing	4	11	
3		219028066	Unknown value	0			Pleasure	3	12	
4		212584000	Moored	0	153	106	Cargo	13	99	6.3
5		636020662	At anchor	0.1	43.9	286	Cargo	23	149	6.3
6		219006116	Unknown value	0	0		Fishing			
7		246539000	At anchor	0	3.8	293	Cargo	16	150	6.8
8		210307000	Moored	0	285.1	225	Cargo	16	90	5.2
9		219003138	Unknown value	0	0		Fishing	4	10	
10		219027820	Unknown value	0.1			Sailing	4	13	

Fig 4.1 snapshot of original dataset

4.1.2 Variables overview

An unrefined and unprocessed, raw AIS dataset consists of a large number of Dynamic, Static and Voyage related information with several fields and millions of rows. It is difficult for a normal computer process such raw information. Hence, after considering a list of databases, a n already processed small section of a bigger database was considered. The fields or the variables included in the data set can be further elaborated and explained as:

Static Information:

1. mmsi- Unique 9-digit identification number
2. shiptype- Type of vessel (tanker, cargo, fishing etc)
3. width – Width of vessel
4. length- Length off vessel
5. draught- the vertical distance between the waterline and the bottom of the hull

Dynamic Information:

6 navigational status - instant situation of a vessel (navigating, mooored etc)
7 sog – speed over ground (usually between 0 to 102 knots)
8 cog – Course over ground 9 0 to 360 degrees)
9 heading - The heading of a vehicle or vessel is its current direction of travel.

All the information mentioned above are required for the analysis. Hence only the column with Sr no is dropped as could create confusion during the interpretation. The variables can be further classified into two categories Numerical and Categorical variables:

- Numerical- mmsi, sog, cog, heading, width, length & draught
- Categorical- navigationalstatus & shiptype

Our main goal from this analysis is to predict the type of ships which are categorical in nature. As mentioned earlier in the literature, there are three types of Machine Learning Problems Supervised, Unsupervised and Reinforcement learning. In our dataset used we have all the values and variables labelled and in supervised machine learning, the model discovers how the labelled input and output data are related to one another. Hence this is a supervised learning problem.

4.2 Exploration of data

4.2.1 Descriptive statistics

The descriptive statistics or the summary statistics in python are obtained by the function describe. Descriptive statistics are brief informative coefficients that summarise a specific data collection. The statistics are categorized into measures of central tendency and measures of variability (Spread). The mean, median, and mode are examples of measurements of central tendency, whereas standard deviation, variance, minimum and maximum variables etc are examples of measures of variability. Central tendency focuses on the centre point of the variables whereas the measure of variability focuses on the shape and spread of data. Below mentioned list depicts all the important functions under descriptive statistics in python pandas. If the database is heterogeneous in nature, then generic operations don’t work with all functions.

Sr.No.	Function	Description
1	count()	Number of non-null observations
2	sum()	Sum of values
3	mean()	Mean of Values
4	median()	Median of Values
5	mode()	Mode of values
6	std()	Standard Deviation of numerical columns
7	min()	Minimum Value
8	max()	Maximum Value
9	abs()	Absolute Value
10	prod()	Product of Values
11	cumsum()	Cumulative Sum
12	cumprod()	Cumulative Product

Table 4.1 important functions of python pandas

Functions like sum() and cumsum() work perfectly on both character (or) string and numeric data elements. But, functions like abs(), cumprod() create an exception when the data consists of character or string data as these operations cannot be executed. The standard deviation is used to calculate how far the data is from the mean. There are in total three types to describe a data, univariate, bivariate and multi variate analysis. Univariate analysis is conducted when there is need to find out spread of single variable. Bivariate and Multi variate analysis are conducted when comparing with two or multi variables. The distribution of numerical variables in our dataset is shown below in the table

4.2.2Distribution of values for Numerical Columns

	Mmsi	Sog	Cog	Heading	Width	Length	draught
Count	358,351.00	357,893.00	355,182.00	337,737.00	354,640.00	354,608.00	332,808.00
Mean	293,967,827.62	12.12	189.06	190.08	19.95	124.97	6.57
Std dev	121,386,631.12	9.36	107.59	107.11	10.81	71.27	2.93
Min	9,112,856.00	0.00	0.00	0.00	1.00	2.00	0.40
25%	219,578,000.00	9.20	116.30	120.00	12.00	83.00	4.60
50%	248,659,000.00	11.30	168.70	170.00	17.00	115.00	6.10
75%	304,665,000.00	13.30	300.18	303.00	28.00	181.00	7.90
Max	992,195,011.00	214.00	359.90	507.00	78.00	690.00	25.50

Table 4.2: Statistics of numerical values of data

The above table provides the statistics of the numerical variables in the dataset. The information of each can further be simplified and explained by visualising the data into graphs

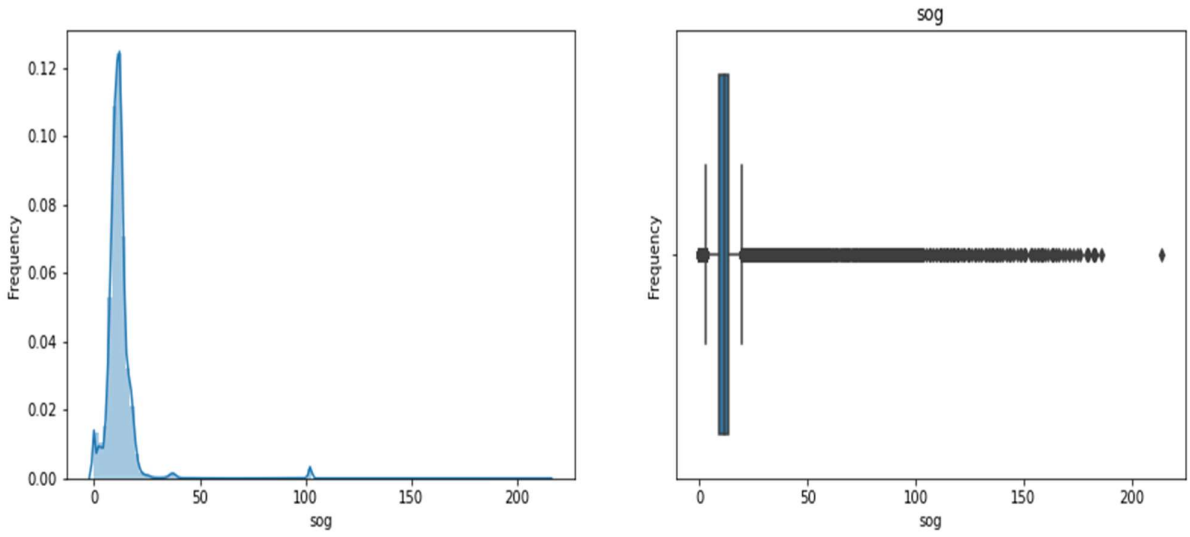

Fig 4.2: Graphical representation of values of speed over ground

The above graphs define the distribution of Speed over ground of ships sailing in the Kattegat region. The average speed of ship is found out to be 12.12 knots with lowest being 00 and highest being 214 knots which seems to be an exceptional case. 75% of the vessels have a speed of approximately 13 knots. The standard deviation of the variable (9.36) is on the lower side and hence the values are not evenly distributed and positively skewed near the mean. From the box plot, it can be inferred that there are several outliers present.

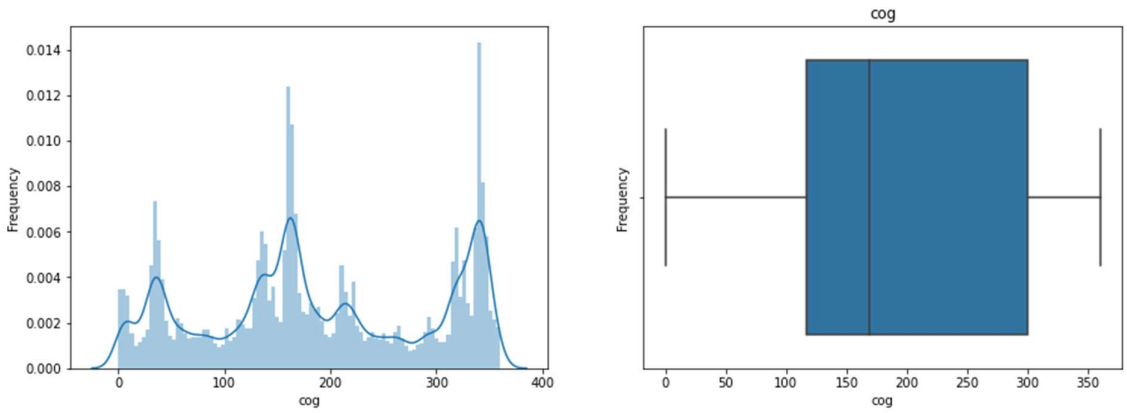

Fig 4.3 Graphical representation of values of course over ground

The course is actual direction of the vessel and is often impacted by strong winds of the ocean. Hence, it is spread between o to 360 degrees, 0.1 displaying north direction, the std deviation is high and hence the values are widely spread far from the average directions. According to the charts average amount of ships head towards the south direction. The box plots suggest that there are no outliers in the data.

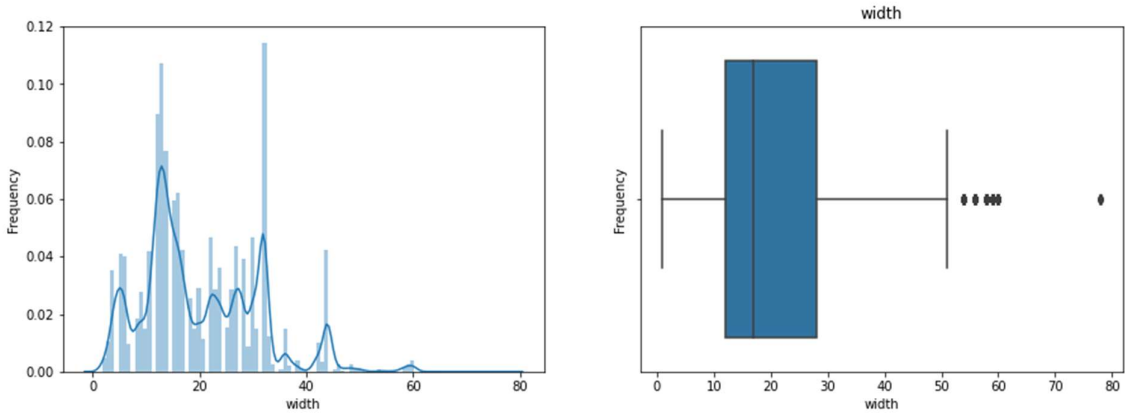

Fig 4.4: Graphical representation of values of width of vessels

The above graph is the distribution of vessels according to their widths. The standard deviation is low and hence the values are not evenly spread. We will focus on the dimensions according to the types of ships in the next coming sections. The box plots suggest that there are certain number of outliers in the data.

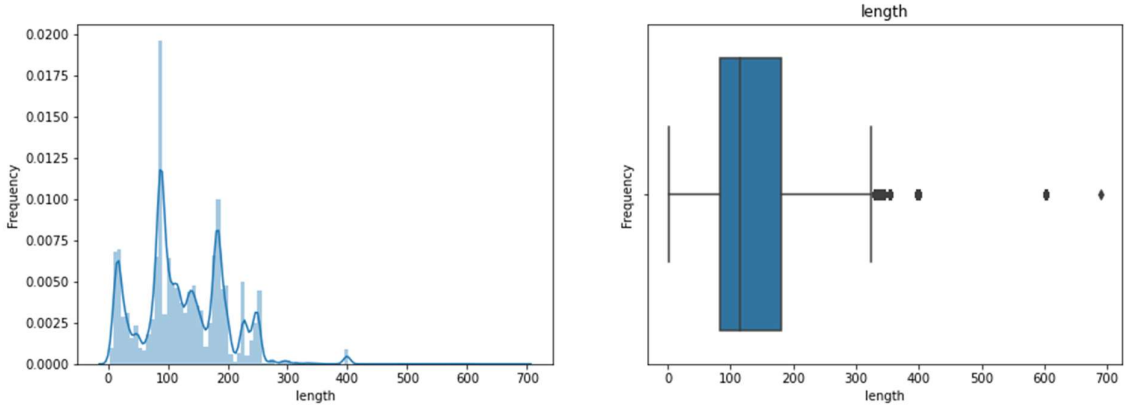

Fig 4.5 Graphical representation of values of length of vessels

The above graph shows the univariate analysis of the variable, length of ships. From the descriptive statistics it is noticed that the std deviation of the variable is on the higher side and as there are several categories of vessels flowing in the region, the variables are spread away from the average value. The box- plot also depicts that there are outliers present in the data.

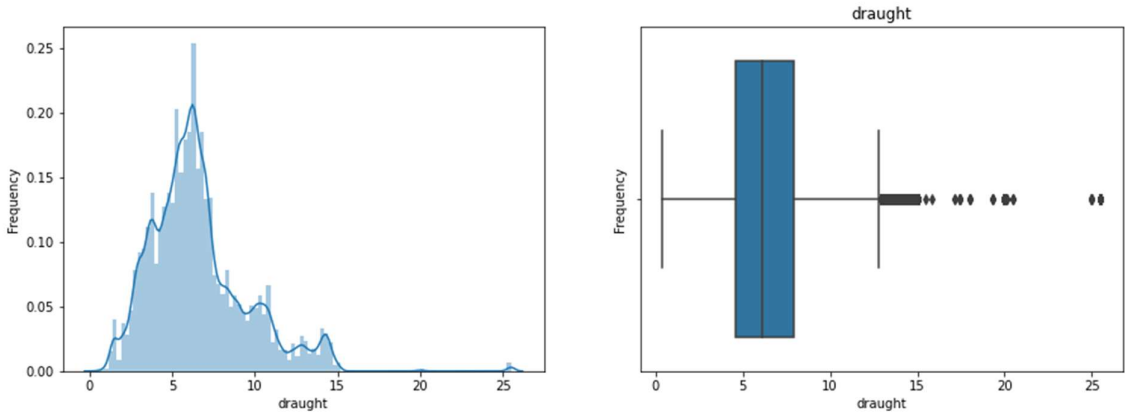

Fig 4.6 Graphical representation of values of draught of vessels

According to the graphs above, it can be inferred that the standard deviations is very low and large number of variables are not evenly spread and thus majority of vessels have a draught size of 7.90. Significant number of outliers are visible in the box plot.

4.2.3 Distribution of values for categorical values

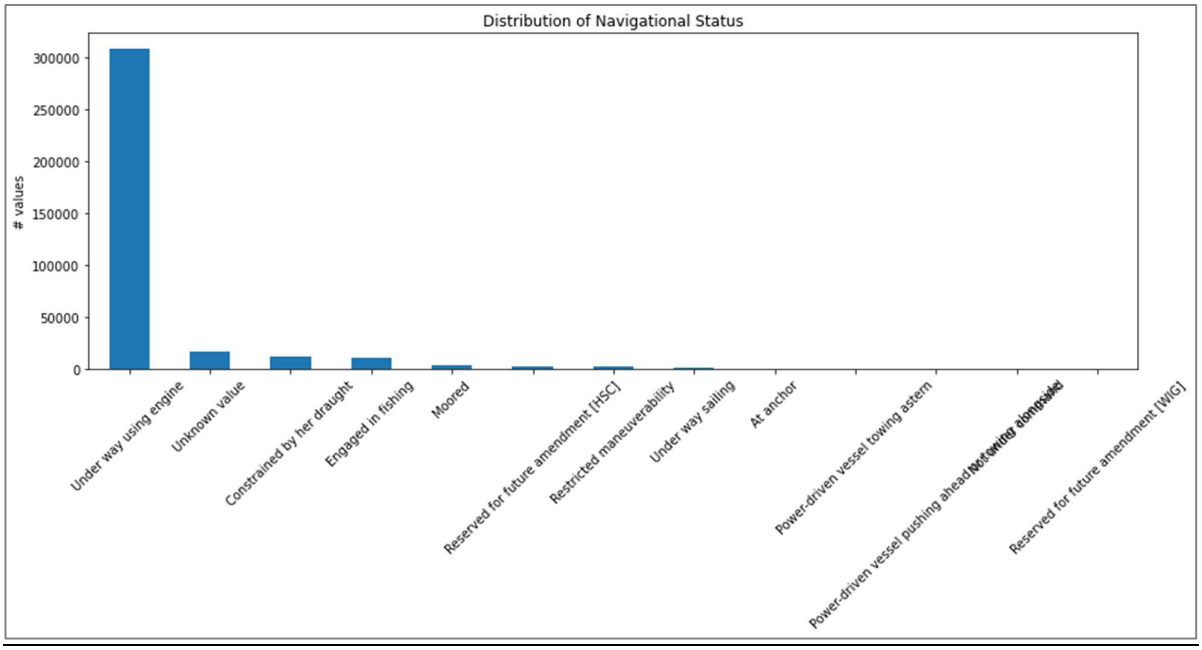

Fig 4.7 Graphical representation of Navigational status of vessels

From the bar chart we can see that majority of ships in our dataset are under way using engine at 86%, while 5% have an unknown navigation status. We also see that there are no ships at anchor or under no command, as well as none under way sailing, telling us all ships in the dataset are currently moving/engaged

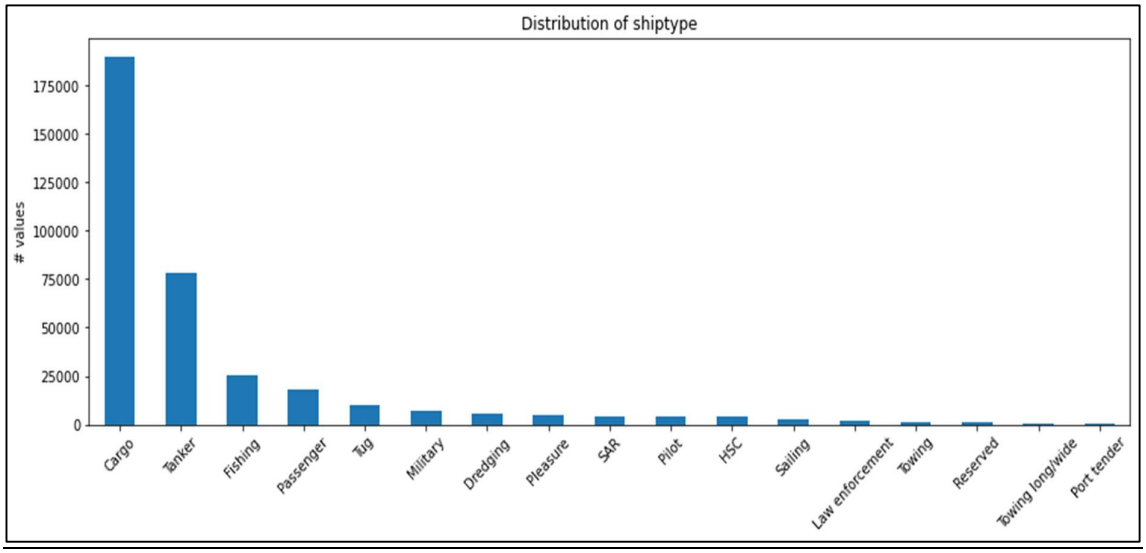

Fig 4.8 Graphical Representation of distribution of ship types

The bar chart displays the types of ships flowing in the waters. As mentioned in most of the ships flowing in the Kattegat region are cargo ships and tankers comprising 53% and 22% of the total population.

4.2.4 Understanding key relationships between variables

Our aim is to predict the ship type by understanding the patterns in the data. Hence, the variable shiptype in the data is our dependent variable. The seaborn heatmap is used to visualize the relation between the target and the other independent variable.

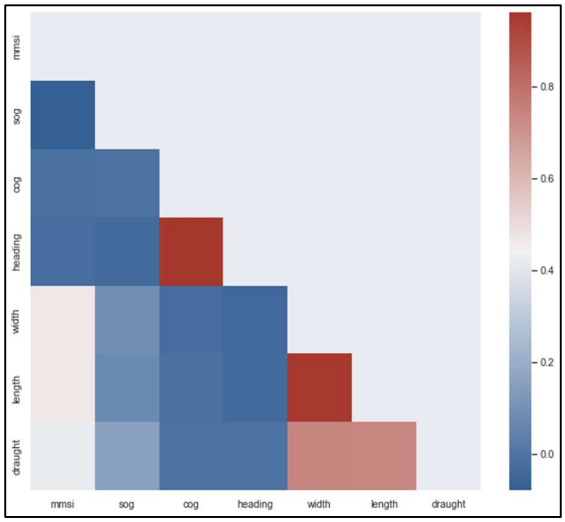

Fig 4.9 correlation matrix

Heading is highly related to cog (96% correlation). Draught on the other hand has 74% correlation with width and length of the ship. Further analysis states that there is clear distinction between shiptype and length and width. Hence, both of the variables are strong predictors in the model to predict the type of ships The average dimensions of vessels according to the vessel types are as follows:

	shiptype	width	length
11	SAR	2.88	9.20
8	Pleasure	3.75	11.30
12	Sailing	4.06	13.66
9	Port tender	4.78	14.00
7	Pilot	5.10	16.79
4	Law enforcement	5.94	24.22
2	Fishing	6.03	22.37
14	Towing	7.45	22.82
15	Towing long/wide	9.00	28.33
5	Military	9.29	50.75
16	Tug	10.58	35.23
10	Reserved	11.92	44.08
1	Dredging	12.38	61.28
3	HSC	13.20	45.68
6	Passenger	16.75	90.51
0	Cargo	21.57	144.02
13	Tanker	28.75	174.91

Table 4.3 Average dimensions of ship types in Kattegat region

4.3 Missing Values and outliers information:

Handling missing values is critical since it can have a negative impact on the findings obtained throughout the research's modelling phase. When analysing the dataset, it was found that Almost 7% of the ships do not have draught values and approximately 6% of the data do have heading values in the data. This gap of data is significantly low and will not affect the model can be ignored. But it was found that there were considerable amount of outliers present in the data. The column speed and draught had 2910 and 567 values abnormally spread across the sample.

4.4 Feature Engineering

As inferred by the heatmap, it is visible that Cog and heading variables are highly correlated with each other. Both variables represent a direction between 0 to 360 degrees. Hence, for better analysis and accurate predictions these both variables were combined into one thereby forming a new variable named “waypoint” which is further divided into 8 directions namely

NNE','ENE','ESE','SSE','SSW','WSW','WNW','NNW. Further, the vessels sailing at speed of below 5.5 knots and with no route information were tagged as “fix”. The missing values in the variable waypoint were handled using median and mode for numerical and categorical variables respectively. The speed of vessels depends on the ship type and hence new variable speed is used to handle the missing values of speed over ground. As mentioned earlier in the literature, length and width are strong predictors for predicting ship type and for example, A tanker sailing in Kattegat region has average length and width of 28.75 and 174.91 respectively. Therefore, a new variable “dimension” is created using the product of length and width to enhance the accuracy of prediction models.

4.5 Creation of dummy variables

Generally, machine learning algorithms are unable to directly act on label data. They demand that all input and output variables be integers or numbers. With help of integer encoding, the algorithms easily predict the natural order of relationship but for category variables where no such relationship exists, the integer coding is not sufficient. On the other hand, if the model is allowed to run assuming a natural order then the model might result in poor performance. Therefore, for adequate accuracy of this model, the variables navigational status and waypoint which are categorical in nature were encoded using the one-hot encoding method. With one-hot, we create a new category column (dummy variables) for each categorical value and assign it a binary value of 1 or 0 and prepare the data for better predictions.

4.6 Splitting the AIS Dataset

It is common practise in machine learning to train and test the model by splitting the data into two distinct sets. These two sets are known as the training and testing sets. The training set, as the name implies, is used to train the model, while the testing set is used to test the model's accuracy. Hence, for better evaluation of our model, our data frame is split into the most common split ratio viz. 80:20 for training and testing set respectively

4.7 Scaling the data

Before fitting the model, it is necessary to scale the data. Because some machine learning algorithms may be dominated by or skewed by input variables with exceptionally big values when compared to the other input variables. As a result, the algorithms tend to disregard the variables with smaller values and focus instead on those with large values. One method of data scaling is called standardization in which we calculate each variable’s mean and standard

deviation and use these values to scale the values such that they have a mean of zero and standard deviation of one, the result is known as a standard normal probability distribution.

But in our data frame the variables speed and draught consist of several outliers which might skew mean and standard deviations in the probability distribution and make the scaling process difficult. Therefore, one method for standardising input variables known as Robust Scaling was applied on the training and the testing set in the analysis. In Robust scaling the outliers are ignored from the calculation of the mean and standard deviation, then scaled with the variables using the values obtained. The resulting variable has zero mean, median and standard deviation as one.

Chapter 5 Model Development

In machine learning, classification consists of two steps: learning and prediction. The model is developed in the learning step using given training data. The model is used in the prediction step to predict the response for given data. In this section, we will discuss about the Machine learning algorithms used to develop the model and perform the analysis. As the research intends to deal with a classification problem, initially a baseline model is created followed by Advanced Machine Learning and Deep Learning Models such as Decision tree, Random Forest, Light BGM classifier and linear SVM classifier. The detailed results of the model are further explained in the following section

5.1 Description of models

1. **Baseline Model**- In a machine learning problem, a baseline model is primarily a simple model that serves as a reference or a benchmark for the trained models. Its primary purpose is to contextualise trained model results and provide better insights of the data. In our study we built a baseline model to determine the most common class amongst the data set which was found out to be cargo ships. The model returned a 54 % accuracy, acting as a reference to other algorithms

2. **Decision Tree Classifier** - The decision tree algorithm, unlike other supervised learning algorithms, can also be used to solve problems involving regression and classification. The main aim of the decision tree algorithm is to predict the class of variable and create a training model by learning simple rules or patterns inferred from the data. It is a tree-structured classifier in which internal nodes represent dataset features, branches signify decision rules, and each leaf node depicts the output. The concept of decision tree is easily explained in the figure below

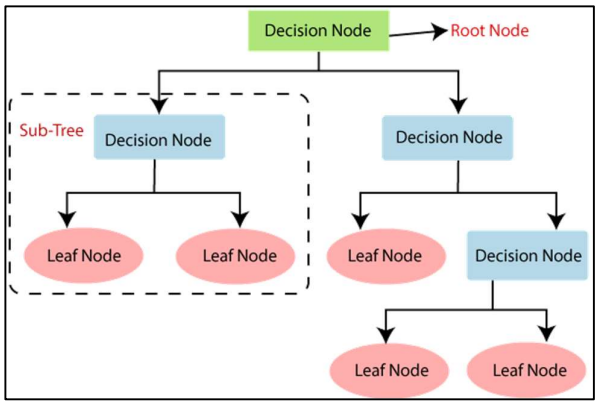

Fig 5.1 Decision tree algorithm

3. **Random Forest Classifier**: A random forest estimator uses averages to improve the prediction accuracy and thus controls overfitting by fitting a number of decision tree classifiers on random sub-samples of the dataset. Because of the large number of decision trees involved in the process, random forests are regarded as a highly accurate and robust method. The algorithm does not suffer from the overfitting problem which is due to the cancellation of biases by taking an average of all the predictions. In our analysis it was found amongst the other algorithms, the Random Forest algorithm provides the best predictive performance and highest accuracy of approximately 98% on the test set.

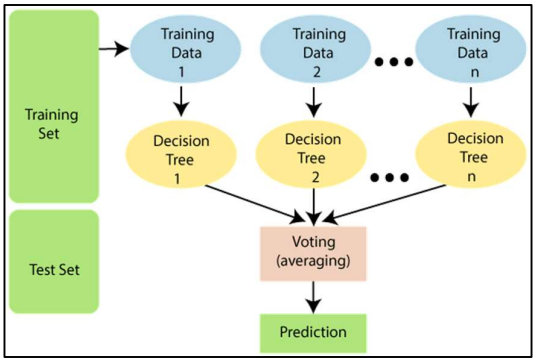

Fig 5.2 Random Forest classifier)

4. **LightGBM Classifier-** One of the powerful algorithms used in complex data science problems is LightGBM Classifier. It is a gradient boosting framework that employs a tree-based learning algorithm. It is a relatively new concept and with the growing amount of data it becomes difficult for the traditional algorithms to give faster results. Unlike the other methods, the classifier prefers leaf- wise growth of nodes as depicted in the figure below. Because of its high speed, Light GBM is prefixed with 'Light.' It can handle large amounts of data while using less memory. Another reason Light GBM

is popular is that it emphasises the accuracy of results. But, there are some limitations in using LightGBM on small datasets. It is prone to overfitting and can easily overfit the limited data. The parameter tuning of the model is filled with complexities as this model covers more than 100 parameters.

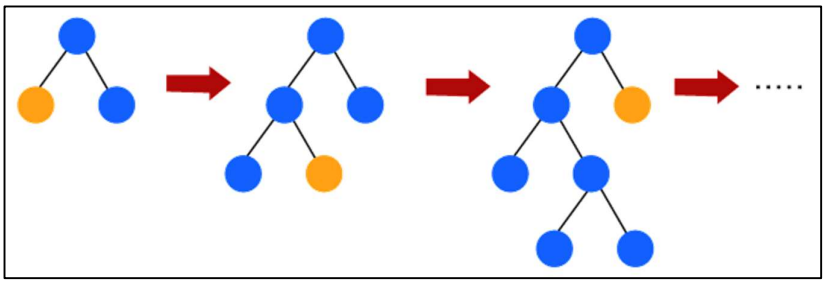

Fig 5.3 LightGBM Classifier

5. **Linear SVM Classifier-** SVM Classifier is a linear model that can solve linear and nonlinear problems and is useful for wide range of practical application. The concept of SVM is simple yet robust, the algorithm draws a line or a hyperplane that divides the data into class. The SVM algorithm's goal is to find the best line for categorising n-dimensional space so that one can easily place new data points in the correct category in the future. This best decision boundary is referred to as a hyperplane. In simple words SVM chooses extreme vectors in the set to create a hyperplane. These extreme vectors are called as support vectors thereby validating the name Linear support vector machine. The figure below explains the concept.

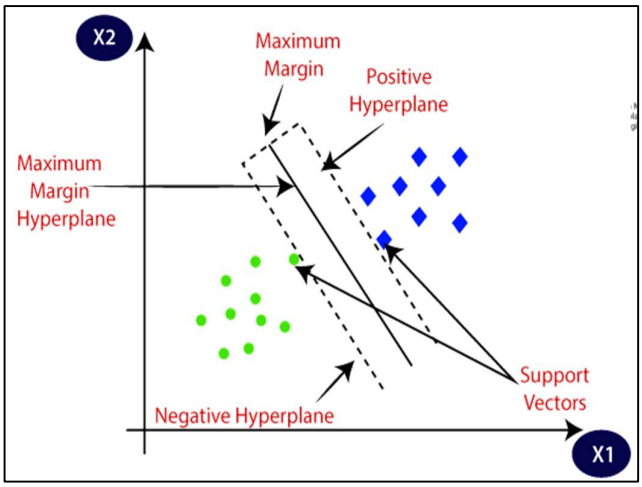

Fig 5.4 Linear SVM Classifier

5.2 Hyperparameter tuning

After creation of a Machine Learning model, The next stage involves a process to choose the best machine learning technique with highest accuracy and tuning it according to the training model's parameters. A model's parameters are the variables that ML techniques use to adjust the data. As studied in the previous section, a deep learning model consists of several nodes throughout the network. These nodes contain a performed operation on the data throughout the network. The customization of number of hidden layer of nodes between input and output layer is not directly related to the training data, these are called hyperparameters which generally are constant during the operation. This tuning of hyperparameters to improve the accuracy and govern the training process is known as hyperparameter tuning.

There are different optimization algorithms which can be used for determining the best hyper parameters, two of the most common methods are Random search and grid search. In our ML Problem we used grid search algorithm to tune the training results. In grid search, a search space is defined with a grid of hyperparameters, and each point in the grid is evaluated for the given vector using cross validation. The detailed results are discussed in the next section

Chapter 6 Model Results and Critical Evaluation

In this part, we will evaluate our predictive models' accuracy and performance on the Automatic Identification system's dataset. For modelling exercise, data was split into train and test sets. Robust scaling was applied to numerical fields and one hot encoding was done for categorical ones. Different ML algorithms were tested, and Random Forest was chosen as the final model based on its superior accuracy over other models. Further we used the Gridsearch cv to determine the best hyperparameters and tuned accordingly to achieve the best results

6.1 Relative performance metrics

The most crucial part of any ML project is to evaluate its performance. Accuracy score, performance metrics, precision, recall, f1 scores are some of the matrices used to evaluate the model.

6.1.1 Accuracy Scores

Accuracy simply reflects how frequently the classifier predicts correctly. Accuracy is calculated as dividing the ratio of true predictions by total predictions. The accuracy scores of all the classifiers used in the analysis are given below. From the table, the Baseline model gives an accuracy score of 54% Random Forest classifier and decision tree classifier perform the best resulting 98% accuracy.

ML Model	Accuracy Scores
Baseline method	54%
Decision Tree classifier	98%
Random Forest Classifier	98%
LightGBM Classifier	68%
Linear SVM Classifier	68%

Table 6.1 Accuracy scores of predictive models

When any model gives such exceptional results, it is not necessary that the model is performing very good. Accuracy is useful when the target is well balanced. But in our cases the target has several values and is not balanced. The model on the other hand resulted in almost similar accuracy scores. Hence, we considered the other performance matrices in addition to accuracy.

6.1.2Precision, Recall and F-1 Scores

From the results and evaluation metrics we see that all models are an improvement from the baseline model. However, we observe very low predictability from the LightGBM and SVM models who have macro-F-score of 0.25 and 0.30 respectively. This indicates that there is a higher prediction of certain types of cargo ships while others are rarely classified which is seen in the very low macro recall scores, indicating the average recall of all ship types.

On the contrary we observe Decision Trees and Random Forest to produce significantly different results with both giving a macro-F-score of 0.91 indicating high predictive power. The recall of each type of ship is similar at 0.91, however we see that the precision of random forest is slightly better than decision trees, with a macro score of 0.92 compared to 0.90 of decision trees. This tells us that a higher no. of records classifying a particular cargo ship turned out to be true, indicating slightly higher predictive power. Hence, we can conclude that Random Forest classifier is the best classification model for our dataset and will be applied to classify the different cargo ship types for each record.

ML Model	Precision	Recall	F-1 Score
Baseline method	0.03	0.06	0.04
Decision Tree classifier	0.90	0.91	0.91
Random Forest Classifier	0.92	0.91	0.91
LightGBM Classifier	0.25	0.27	0.25
Linear SVM Classifier	0.41	0.30	0.30

Table 6.2 Evaluation matrices of predictive models

We further improved RF models' predictions using Hyperparameter tuning. We defined a grid and specified discrete values using the GridsearchCV which resulted in 0.91 F-1 Macro score and used it to tune the whole model. At the end of the run, it was found that the RF classifier varied greatly ranging from a mean test F-1 score between 0.45 to 0.91. Also, it was found that increasing the value of max depth results in better performance. Decreasing the sample split might improve the performance but too little value can lead to overfitting of the model.

Further, a new RF model was created using the best parameters resulting in improved performance with 0.92 macro-F-1 Score and 0.91 recall rate which indicates that there is more

accuracy in prediction of categories of cargo ships. With the table and graph below we assigned scores to the input features that indicates the relative features of each feature while making a prediction. Therefore, it can be inferred that the most important features required to accurately predict the type of vessels flowing in the Kattegat region are length, dimension, and width. The individual scores are mentioned below.

Sr No.	**Features**	**Scores in Percentage (%)**
1	Length	24
2	Dimension	23
3	Width	15
4	Draught	15
5	Speed	6
6	cog	4
7	sog	4

Table 6.3 Importance of features for the model

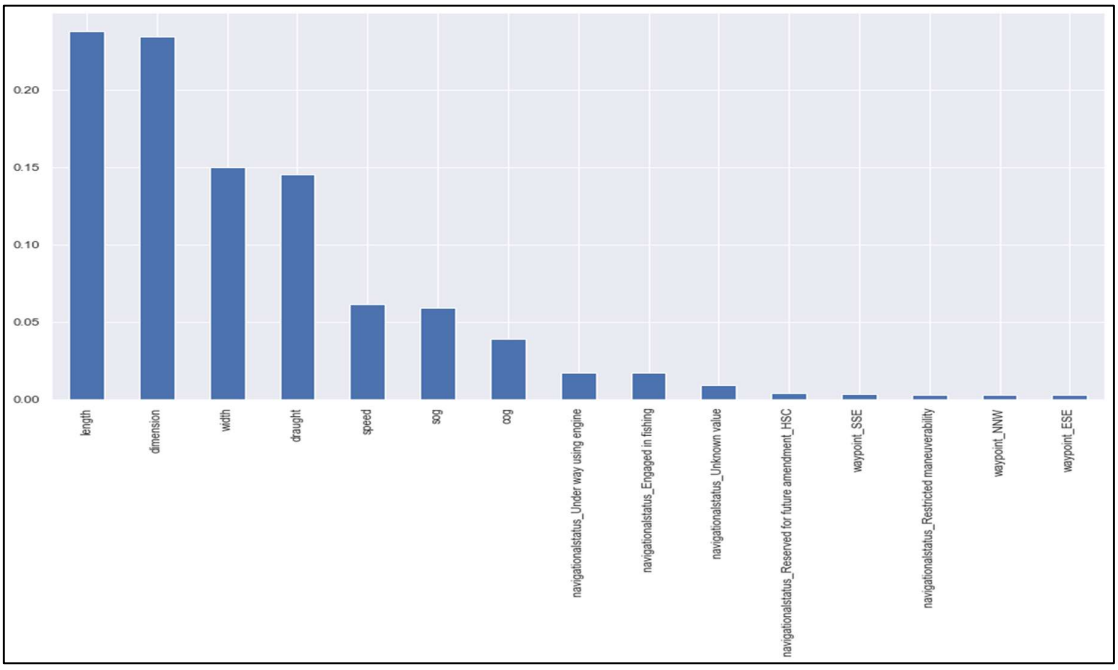

Fig 6.1 Top features and important variables for predicting vessel types

Thus, we conclude our analysis, we successfully built a predictive model to predict the type of ships flowing in the region of Kattegat strait by collecting and processing AIS dataset of 3 months of that region. From the model, we validate an information that majority of the ships

flowing in the waters are cargo vessels and tanker ships. Length, width, and dimensions are the most important parameters which are utilized to predict the type of ship. This model can be used by organizations and stakeholders to train the computers in order to find out the type of vessel using their dimensions. From the results it can also be inferred that the ports in the region of Baltic sea currently do not possess an infrastructure to match the rising demand of shipping operations. Therefore, should strategically prepare contingency plans to avoid maritime incidents in the Kattegat region.

Chapter 7: Conclusion and Future Research

7.1 Conclusion

Here, we reached the conclusion of our study. In this section we summarise with our findings throughout the whole research. Several types of machine learning and Advanced Machine learning algorithms were used in this research to predict the type of vessels flowing in the Baltic region, and we successfully built a model with superior results. Various performance metrics are used to analyse different models' accuracy and performance, as shown in the results section of this research.

The project had four key goals to achieve using the machine learning algorithms. Firstly, Collect historical data and understand how to get insight from the data that is currently available. Secondly, To find methodologies and create a model that can be used to make predictions based on historical AIS data. Thirdly, Determine the model's dependability through testing and comparison. Lastly, To provide a prototype that can be implemented by existing stakeholders for the optimization and safety of the shipping sector.

The motivation to conduct this research was the shipping and maritime challenges in the Kattegat Strait region of Baltic Sea. Another stimulation was to imply the learnt theories and commonly used ML models on the Automatic Identification System Dataset. A huge amount of study and research has been undertaken on shipping and maritime logistics and Application of different ML techniques. Following an analysis of various classification machine learning models, the results show that a Random Forest classifier performs the best for predicting the kind of ship.

The model beats advanced machine learning algorithms, achieving 98% accuracy, 92% precision, and 91% recall. Therefore, we can say that our model outperforms the other techniques and predicts more number of ships accurately. Further this model can be used to accurately determine the type of ships for practical or for research purposes. It can also be inferred that three months of AIS data is sufficient to solve the classification problem.

References

Works Cited

International Institute for Law of the Sea Studies, 2021. *Navigational Regimes of Particular Straits, Baltic Straits(The Oresund and the Belts) case study.* [Online]
Available at: http://iilss.net/navigational-regimes-of-particular-straits-baltic-straitsthe-oresund-and-the-belts-case-study/
[Accessed 30 August 2022].

(Nina), N. H. B., 2019. *Predicting arrival times of container vessels - A machine learning application,* Oldenzaal: University of twente.

A Grimvall, K. l., 2014. *Mapping shipping intensity and routes in the Baltic Sea.,* s.l.: Swedish Institute for the Marine Environment.

Abbas Harati-Mokhtari, A. W. P. B. J. W., 2007. Automatic Identification System (AIS): Data Reliability and Human Error Implications. *The journal of Navigation,* 60(3), pp. 373-389.

Abebe, M. &. S. Y. &. N. Y. &. L. S. &. L. I., 2020. Machine Learning Approaches for Ship Speed Prediction towards Energy Efficient Shipping. *Applied sciences.*

Agnieszka, K., 2018. Traffic Rules and Environmental Conditions in Kattegat and the Sound Regarding Changes Planned for 2020. *Scientific Journal of Gdynia Maritime University,* Volume 107.

Anon., 2018. *Association for advancement of artificial intelligence.* [Online]
Available at: https://www.aaai.org/

An-sing Chen, M. T. L., 2004. Regression neural network for error correction in foreign exchange forecasting and trading. *Computers & Operations Research,* 31(7), pp. 1049-1068.

Atak, A. &., 2021. Machine learning methods for predicting marine port accidents: a case study in container terminal. *Ships offshore structure.*

Bhavsar, P., 2017. *Machine Learning in Transportation Data Analytics,* s.l.: Elsevier Inc.

Chen, Z. &. C. D. &. Z. Y. &. C. X. &. Z. M. &. W. C., 2016. Deep learning for autonomous ship-oriented small ship detection. *Safety Science,* 130(9).

Ching-Jung Ting, K.-C. W. H. C., 2014. Particle swarm optimization algorithm for the berth allocation problem. *Experts systems with applications,* 41(4), pp. 1543-1550.

Christiaan Heij, S. K., 2019. Shipping inspections, detentions, and incidents: an. *Maritime Policy & Management,* 46(7).

claudia caballini, M. D. G. J. M.-. O. S. S., 2020. A combined data mining – optimization approach to manage trucks operations in container terminals with the use of a TAS: Application to an Italian and a Mexican port. *Transportation Research Part E: Logistics and Transportation Review,* Volume 142.

Cong Liu, J. L. X. Z. Z. Z. C. W. Z. L., 2020. AIS data-driven approach to estimate navigable capacity of busy waterways focusing on ships entering and leaving port,. *Ocean Engineering,* Volume 218.

Davies, R., 2020. *Managing sea traffic in the Baltic Sea.* [Online]
Available at: https://www.ship-technology.com/analysis/baltic-sea-shipping-routes/
[Accessed 08 2022].

Davies, R., 2020. *Managing sea traffic in the Baltic Sea.* [Online]
Available at: https://www.ship-technology.com/analysis/baltic-sea-shipping-routes/
[Accessed 30 August 2022].

Digital, M., n.d. *Maritime technology challenges 2030.* [Online]
Available at: https://marine-digital.com/article_maritime_challenges_2030

Dong Yang, L. W. S. W. H. j. K. X. L., 2019. How big data enriches maritime research – a critical review of Automatic Identification System (AIS) data applications. *Transport reviews,* 39(6), pp. 755-773.

ECSA, 2019-24. Sailing ahead: European shipping sets ambitious goals for its next chapter. *Strategic priorities for EU shipping policy 2019–2024,* pp. 3-4.

EMSA, n.d. *COVID-19: impact on the maritime sector in the EU.* [Online]
Available at: https://www.emsa.europa.eu/COVID19

Enmei Tu, G. Z. S. M. L. R. G.-B. H., 2020. Modeling Historical AIS Data For Vessel Path. *ArXiv,* Volume 2001.01592v3.

Ernest Czermanski, M. M. N. K. M. E. O. L. S. M. V. H. W. J. Z. C. C. A. K., 2018. *QUO VADIS Exploring the future of shipping in the Baltic Sea,* s.l.: s.n.

Freitas, A. A., 2002. *Data Mining and Knowledge Discovery with Evolutionary Algorithms.* s.l.:Springer.

Han, C.-H., 2018. *Assessing the Impacts of Port Supply Chain Integration on Port Performance,* Busan: Asian Journal of Shipping and Logistics.

HELCOM, 2018. *Report on shipping accidents in Baltic Sea from 2014 to 2017,* s.l.: s.n.

HELCOM, n.d. *ensuring safe shipping in the Baltic,* finland: s.n.

International Institute for law of Sea Studies, 2021. *Navigational Regimes of Particular Straits, Baltic Straits(The Oresund and the Belts) case study.* [Online]
Available at: http://iilss.net/navigational-regimes-of-particular-straits-baltic-straitsthe-oresund-and-the-belts-case-study/
[Accessed 30 August 2022].

International Telecommunication Union, 2014. *Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime frequency band.* [Online]
Available at: https://www.itu.int/rec/R-REC-M.1371

Interreg Baltic sea region, 2016. *SHIPPING IN THE BALTIC SEA,* s.l.: s.n.

Ioanna Kourounioti, A. P. C. T., 2016. Development of Models Predicting Dwell Time of Import Containers in Port Container Terminals – An Artificial Neural Networks Application. *Transportation Research Procedia,* Volume 14, pp. 243-252.

James C, J. W., 2008. *The Impacts of Globalisation on International Maritime Transport Activity,* Mexico: OECD/ITF.

Johansson, U., 2007. *Obtaining Accurate and Comprehensible Data Mining Models - an evolutionary approach,* Linköping: Department of Computer and Information Science Linköpings universitet.

Kadir Cicekc, U. O. S. I. B., 2019. Evaluating navigational risk of port approach manoeuvrings with expert assessments and machine learning. *Ocean Engineering,* Volume 192, pp. 1-21.

Kantardzic, M., 2011. *Data Mining: Concepts, Models, Methods and Algorithms,* New Jersey: John Wiley & Sons, Inc.

LAM, Y., 2020. *Technology will help maritime transport navigate through the pandemic—and beyond.* [Online]
Available at: https://blogs.worldbank.org/transport/technology-will-help-maritime-transport-navigate-through-pandemic-and-beyond

Lecq, L. v. d., 2021. *Mapping Maritime Risk in the Kattegat Using the Automatic Identfication System,* utretch: UtrechtUniversity.

Li Da Xu, E. L. X. &. L. L., 2018. *Industry 4.0: state of the art and future trends,* s.l.: International Journal of Production Research.

Lutz Kretschmann, M. Z. S. K. T. H., 2019. *MACHINE LEARNING IN MARITIME LOGISTICS,* Hamburg: Fraunhofer CML.

Marielle Christiansen, K. F. B. N. D. R., 2013. Ship routing and scheduling in the new millennium. *European Journal of Operational Research,* 228(31), pp. 467-483.

Nadereh Moini, M. B. S. T. W. L., 2012. Estimating the determinant factors of container dwell times at seaports. *Maritime Economics & Logistics* , may, pp. 162-177.

NATO shipping centre, 2021. *AIS (AUTOMATIC IDENTIFICATION SYSTEM) OVERVIEW.* [Online]
Available at: https://shipping.nato.int/nsc/operations/news/2021/ais-automatic-identification-system-overview
[Accessed 2022].

Pang-Ning Tan, M. S. V. K., 2006. *Introduction to Data Mining,* s.l.: pearson Addison-Wesley.

Pena, B., 2020. *A Review on Applications of Machine Learning in Shipping,* Houston: SNAME Maritime Convention 2020.

Pentti Kujala, L. D. F. G., 2020. Review and analysis of methods for assessing maritime waterway risk based on non-accident critical events detected from AIS data. *Reliability Engineering and safety,* 200(106933).

Pete Chapman (NCR), J. C. (. R. K. (. T. K. (. T. R. (. C. S. (. a. R. W. (., 2000. *CRISP-DM 1.0 - Step-by-step data mining guide,* s.l.: s.n.

Ran Yan, S. W. C. P., 2021. An Artificial Intelligence Model Considering Data Imbalance for Ship Selection in Port State Control Based on Detention Probabilities. *Journal of Computational Science,* Volume 48.

René Taudal Poulsen, H. S., 2020. A swift turnaround? Abating shipping greenhouse gas emissions via port call optimization. *Transportation Research Part D: Transport and Environment,* Volume 86.

Riveiro, M. &. P. G. &. V. M., 2018. Maritime anomaly detection: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.*

Sea Traffic Managmenet , 2019. *Validation Project report,* s.l.: s.n.

Shankar, S. I. P. P. S. a. S. S., 2020. Forecasting container throughput with long short-term memory networks. *Industrial Management & Data Systems,* 120(3), pp. 425-441.

Sinay, 2021. *What is Artificial Intelligence in Smart Port Operations?.* [Online]
Available at: https://sinay.ai/en/what-is-artificial-intelligence-in-smart-port-operations/
[Accessed August 2022].

Siyavash Filom, A. M. A. S. R., 2022. *Applications of Machine Learning methods in port operations - A systematic literature review,* Hamilton: ScienceDirect.

Steven C. Mallam, S. N. A. S., 2020. The human element in future Maritime Operations – perceived impact of autonomous shipping. *Ergonomics,* 63(3), pp. 334-345.

Ties Emmens, C. A. A. A. M. G., 2021. The promises and perils of Automatic Identification System data. *Expert Systems With Applications,* 178(114975).

Trudi Hogg, S. G., 2016. Autonomous merchant vessels: examination of factors that impact the effective implementation of unmanned ships. *Australian Journal of Maritime & Ocean Affairs ,* 8(3), pp. 206-222.

UNCTAD, 2018. *Review maritime Transport.* [Online]
Available at: https://unctad.org/webflyer/review-maritime-transport-2018

UNCTAD, 2019. *DIGITALIZATION IN MARITIME TRANSPORT:ENSURING OPPORTUNITIES FOR DEVELOPMENT,* s.l.: Digital Container Shipping Association.

Wen-Yi-Peng, C.-W., 2009. A comparison of univariate methods for forecasting container throughput volumes. *Mathematical and Computer Modelling,* 50(7-8), pp. 1045-1057.

Wingrove, M., 2020. *Satellite investment boosts cruise ship connectivity.* [Online]
Available at: https://www.rivieramm.com/news-content-hub/news-content-hub/satellite-investment-boosts-cruise-ship-connectivitynbsp-58659
[Accessed September 2022].

Yamin Huang, L. C. P. C. R. R. N. P. v. G., 2020. Ship collision avoidance methods: State-of-the-art. *Safety Science,* Volume 121, pp. 451-473.

Yi Xiao, G. W. K.-C. L. G. Q. K. X. L., 2020. The effectiveness of the New Inspection Regime for Port State Control: Application of the Tokyo MoU. Volume 115.

Zarzuelo, I. d. l. P., 2020. *Industry 4.0 in the port and maritime industry: A literature review.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S2452414X20300480

Appendices

1.1 Importing and Interpreting the use of libraries

Pandas is flexible, fast and easy to use open source data analysis and manipulation tool.
Numpy is python library used to perform mathematical operations on array.
matplotlib and **seaborn** are python libraries used for data visualization. We can visualize data by charts and plots.
sklearn is used for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

In [1]:
```
import pandas as pd
import numpy as np

# Libraries for plotting and visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Libraries for modeling
from lightgbm import LGBMClassifier
from sklearn.preprocessing import RobustScaler
from sklearn.model_selection import train_test_split,cross_val_score,GridSearchCV,RandomizedSearchCV
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

import time
import warnings
warnings.filterwarnings(action='ignore')
```

1.1 Importing the libraries

In [3]:
```
#Get information for the dataset such as datatype and non null counts
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 358351 entries, 0 to 358350
Data columns (total 9 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   mmsi                358351 non-null  int64
 1   navigationalstatus  358351 non-null  object
 2   sog                 357893 non-null  float64
 3   cog                 355182 non-null  float64
 4   heading             337737 non-null  float64
 5   shiptype            358351 non-null  object
 6   width               354640 non-null  float64
 7   length              354608 non-null  float64
 8   draught             332808 non-null  float64
dtypes: float64(6), int64(1), object(2)
memory usage: 24.6+ MB
```

Data has total 9 columns with 7 being numerical and categorical.

Numerical columns: mmsi, sog, cog, heading, width, length & draught
Categorical columns: navigationalstatus & shiptype

1.2 Initial description

2.1. Missing value information

In [4]:
```
pd.options.display.float_format = "{:,.3f}".format

def missing_values_table(df):
    m=df.isnull().sum()
    print(pd.DataFrame({'n_miss' : m[m!=0],'% of total count' : m[m!=0]/len(df)}))
```

```
missing_values_table(df)
```

```
         n_miss  % of total count
sog         458             0.001
cog        3169             0.009
heading   20614             0.058
width      3711             0.010
length     3743             0.010
draught   25543             0.071
```

Column **draught** has ~7.1% and **heading** has ~5.7% missing values.
width and **length** each has around 1% missing values.

2.1 Missing Value Information

2.2. Numerical columns

1. Summary statistics for Numerical columns

In [5]:
```
pd.options.display.float_format = "{:,.2f}".format
num_cols = ['mmsi','sog','cog','heading','width','length','draught']

df[num_cols].describe()
```

Out[5]:

	mmsi	sog	cog	heading	width	length	draught
count	358,351.00	357,893.00	355,182.00	337,737.00	354,640.00	354,608.00	332,808.00
mean	293,967,827.62	12.12	189.06	190.08	19.95	124.97	6.57
std	121,386,631.12	9.36	107.59	107.11	10.81	71.27	2.93
min	9,112,856.00	0.00	0.00	0.00	1.00	2.00	0.40
25%	219,578,000.00	9.20	116.30	120.00	12.00	83.00	4.60
50%	248,659,000.00	11.30	168.70	170.00	17.00	115.00	6.10
75%	304,665,000.00	13.30	300.18	303.00	28.00	181.00	7.90
max	992,195,011.00	214.00	359.90	507.00	78.00	690.00	25.50

2.2 Summary Stats for numerical columns

3. Frequency distribution and Box-Plot

In [7]:
```
col_list = ['sog', 'cog', 'heading','width', 'length', 'draught']
```

```
for col in col_list:
    fig = plt.figure(figsize = (15,5))
    #Histogram
    plt.subplot(1,2,1)
    #Define plot object
    hist = sns.distplot(df.loc[:,col].astype(float), bins = 100)
    #Setting graph title
    #hist.set_title(col)
    hist.set(xlabel = col, ylabel = 'Frequency')
    #Boxplot
    plt.subplot(1,2,2)
    #Define plot object
    box = sns.boxplot(df.loc[:,col].astype(float))
    #Setting graph title
    box.set_title(col)
    box.set(xlabel = col, ylabel = 'Frequency')
    #Showing the plot
    plt.show()
```

Plotting frequency distribution and box plot

2.3. Distribution of values for categorical columns

In [8]:

```
# Distribution of column "navigationalstatus"

print('% Distribution of Navigational Status:\n')
print(df['navigationalstatus'].value_counts(1))

print('\n\n')
plt.figure(figsize=(15,5))
df['navigationalstatus'].value_counts().plot.bar()
plt.title('Distribution of Navigational Status')
plt.ylabel('# values')
plt.xticks(rotation = 45)
plt.show()
```

```
% Distribution of Navigational Status:

Under way using engine                                0.86
Unknown value                                         0.05
Constrained by her draught                            0.03
Engaged in fishing                                    0.03
Moored                                                0.01
Reserved for future amendment [HSC]                   0.01
Restricted maneuverability                            0.01
Under way sailing                                     0.00
At anchor                                             0.00
Power-driven vessel towing astern                     0.00
Power-driven vessel pushing ahead or towing alongside 0.00
Not under command                                     0.00
Reserved for future amendment [WIG]                   0.00
Name: navigationalstatus, dtype: float64
```

2.3 Distribution of Categorical Values

2.4. Outlier detection using Box-Plot method

In [10]:

```
def thresholds(col, data, d, u):
    q3=data[col].quantile(u)
    q1=data[col].quantile(d)
    down=q1-(q3-q1)*1.5
    up=q1+(q3-q1)*1.5
    return down, up

def check_outliers(col, data, d=0.25, u=0.75, plot=False):
    down, up = thresholds(col, data, d, u)
    ind = data[(data[col] < down) | (data[col] > up)].index
    if plot:
        sns.boxplot(x=col, data=data)
        plt.show()
    if len(ind)!= 0:
        print(f"\n Number of outliers for '{col}' : {len(ind)}")
        return col

for col in num_cols:
    check_outliers(col, df, 0.01, 0.99) # we set thresholds at 0.01 and 0.99
```

```
Number of outliers for 'mmsi' : 24

Number of outliers for 'sog' : 2910

Number of outliers for 'width' : 9

Number of outliers for 'length' : 47

Number of outliers for 'draught' : 567
```

2.4 Outlier Detection

2.5. Correlation Matrix between numerical columns

[11]:

```
# correlation matrix dataframe
df_corr = df[num_cols].corr()

### Heat map of correlation matrix
sns.set_theme(style="darkgrid")

mask = np.triu(np.ones_like(df_corr))
f, ax = plt.subplots(figsize=(11, 9))
cmap = sns.diverging_palette(250, 15, s=75, l=40, center="light", as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(df_corr, mask=mask, cmap=cmap, square=True)
plt.show()
```

2.5 Plotting Correlation Matrix

2.6 Average length and width by shiptype

[12]:

```
df[['mmsi','shiptype','length','width']].drop_duplicates(subset='mmsi')\
                                        .groupby(['shiptype'])\
                                        .agg({'width':'mean','length':'mean'})\
                                        .reset_index()\
                                        .sort_values('width')
```

2.6 Average length and width of shiptypes

3.1. Creating new features

1. 'cog' and 'heading' variables are very similar. And they also have a high correlation of 96% between them. So we can combine these two as one.
2. Dividing the 360-degree route into 8 regions.
3. the ships with less than 5.5kts speed and no route information were tagged as 'FIX'.
4. Vessels' speed depends on ship type mostly. We fill the missings according to 'sog' and 'route' variables. Then assigned as a new variable "speed"
5. new variables dimension = widht*length

[13]:

```
# First, the filling was made according to those in the 'heading' but not in the 'cog'.
df['cog'] = np.where(df['cog'].isnull(), df['heading'], df['cog'])

# Secondly, we divided the 360-degree route into 8 regions.
rot= [-1, 45, 90, 135, 180, 225, 270, 315, 360]
df['waypoint'] = pd.cut(df['cog'], rot, labels=['NNE','ENE','ESE','SSE','SSW','WSW','WNW','NNW'])

# Finally, the ships with less than 5.5kts speed and no route information were tagged as 'FIX'.
df['waypoint'] = np.where((df['sog']<5.5) & (df['waypoint'].isnull()), 'FIX', df['waypoint'])

# Filling the missings according to 'sog' and 'cog' variables. Then assigned as a new variable "speed".
df['speed'] = df["sog"].fillna(df.groupby(['shiptype', 'waypoint'])['sog'].transform('mean'))

# dimension = lenght*width
df['dimension'] = df['width'] * df['length']
```

3.1 Feature engineering

3.2. Dropping features not required for modeling and deduping

[14]:

```
# columns to drop
drop_cols = ['mmsi','heading']

# new df_m for modeling data only
df_m = df.drop(drop_cols, axis=1).drop_duplicates()

print('shape modeling data after dropping columns and deduping', df_m.shape)
```

```
shape modeling data after dropping columns and deduping (348379, 10)
```

3.2 &3.3 Dropping irrelevant features

4. OHE - one-hot encoding for categorical features

Only "navigationalstatus" and "waypoint" will be encoded and shiptype since shiptype is the model target

```
[6]:
# One hot encoder function;
def one_hot_encoder(df, cat_cols, drop_first=True):
    dataframe = pd.get_dummies(df, columns=cat_cols, drop_first=drop_first)
    return dataframe
```

```
[7]:
df_m = one_hot_encoder(df_m, cat_cols=['waypoint','navigationalstatus'], drop_first=True)
df_m
```

3 One hot encoding for categorical variables

5. Data-Splitting

Splitting the data into train and test for modelling

Train 80% and Test 20%

5.1 Feature and Target split

```
[19]:
target = 'shiptype'

X = df_m.drop(target, axis=1)
y = df_m[target]
```

5.2 Train-Test split

```
[20]:
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=7)
```

5, 5.1, 5.2 Data splitting

5.3 Scaling the features - Robust Scaler

'RobustScaler' is used because it is robust to outliers

```
[21]:
# RobustScaler object - fitting on train
scaler = RobustScaler()
scaler.fit(X_train)

# transforming X_train
X_train = pd.DataFrame(scaler.transform(X_train), columns=X_train.columns)

# transforming X_test
X_test = pd.DataFrame(scaler.transform(X_test), columns=X_test.columns)
```

```
[ ]:
```

5.3Robust scaler

```
[22]:
## Function to plot feature importance
def show_feat_imp(model, X, n_feats=15):
    """
    model: model object
    X: feature dataset on which prediction is to be done. Ex: X_test or X_train
    n_feats: number of top features to display
    """
    feat_imp = pd.Series(model.feature_importances_, index = X.columns).sort_values(ascending = False)
    display(feat_imp.head(n_feats))
    feat_imp.head(n_feats).plot.bar(x='features', y='feat_imp', figsize=(15,8), align="center")
```

6.0 Modelling

```
In [25]: # performance of baseline model on test
         print(classification_report(y_test,prediction))
```

	precision	recall	f1-score	support
Cargo	0.54	1.00	0.70	37524
Dredging	0.00	0.00	0.00	1062
Fishing	0.00	0.00	0.00	4986
HSC	0.00	0.00	0.00	709
Law enforcement	0.00	0.00	0.00	320
Military	0.00	0.00	0.00	1416
Passenger	0.00	0.00	0.00	3298
Pilot	0.00	0.00	0.00	789
Pleasure	0.00	0.00	0.00	625
Port tender	0.00	0.00	0.00	57
Reserved	0.00	0.00	0.00	144
SAR	0.00	0.00	0.00	709
Sailing	0.00	0.00	0.00	370
Tanker	0.00	0.00	0.00	15452
Towing	0.00	0.00	0.00	208
Towing long/wide	0.00	0.00	0.00	122
Tug	0.00	0.00	0.00	1885
accuracy			0.54	69676
macro avg	0.03	0.06	0.04	69676
weighted avg	0.29	0.54	0.38	69676

The accuracy of baseline model is at 54%. It can definitely be improved.

6.1 Baseline model results

	precision	recall	f1-score	support
Cargo	0.99	0.99	0.99	37524
Dredging	0.97	0.97	0.97	1062
Fishing	0.96	0.96	0.96	4986
HSC	0.99	0.98	0.99	709
Law enforcement	0.98	0.98	0.98	320
Military	0.99	0.99	0.99	1416
Passenger	0.99	0.99	0.99	3298
Pilot	0.96	0.98	0.97	789
Pleasure	0.65	0.62	0.64	625
Port tender	0.66	0.72	0.69	57
Reserved	0.92	0.90	0.91	144
SAR	0.89	0.90	0.89	709
Sailing	0.59	0.61	0.60	370
Tanker	0.98	0.98	0.98	15452
Towing	0.90	0.91	0.91	208
Towing long/wide	0.96	0.95	0.95	122
Tug	0.99	0.99	0.99	1885
accuracy			0.98	69676
macro avg	0.90	0.91	0.91	69676
weighted avg	0.98	0.98	0.98	69676

Accuracy of the model is **98%** which is quite good for a simple decision tree.

6.2 Decision tree classifier results

	precision	recall	f1-score	support
Cargo	0.99	0.99	0.99	37524
Dredging	0.98	0.96	0.97	1062
Fishing	0.96	0.97	0.97	4986
HSC	1.00	0.99	0.99	709
Law enforcement	0.99	0.98	0.99	320
Military	0.99	0.98	0.98	1416
Passenger	0.99	0.99	0.99	3298
Pilot	0.97	0.98	0.97	789
Pleasure	0.68	0.65	0.66	625
Port tender	0.76	0.74	0.75	57
Reserved	0.90	0.90	0.90	144
SAR	0.92	0.90	0.91	709
Sailing	0.62	0.64	0.62	370
Tanker	0.99	0.97	0.98	15452
Towing	0.91	0.91	0.91	208
Towing long/wide	0.97	0.91	0.94	122
Tug	0.99	0.99	0.99	1885
accuracy			0.98	69676
macro avg	0.92	0.91	0.91	69676
weighted avg	0.98	0.98	0.98	69676

Random forest has similar performance like decision tree with **98%** accuracy on test set.

6.3 Initial random forest classifier

	precision	recall	f1-score	support
Cargo	0.83	0.83	0.83	37524
Dredging	0.11	0.05	0.07	1062
Fishing	0.58	0.69	0.63	4986
HSC	0.00	0.00	0.00	709
Law enforcement	0.18	0.25	0.21	320
Military	0.36	0.34	0.35	1416
Passenger	0.57	0.68	0.62	3298
Pilot	0.28	0.40	0.33	789
Pleasure	0.13	0.21	0.16	625
Port tender	0.00	0.00	0.00	57
Reserved	0.00	0.00	0.00	144
SAR	0.07	0.13	0.09	709
Sailing	0.09	0.09	0.09	370
Tanker	0.78	0.60	0.68	15452
Towing	0.03	0.16	0.05	208
Towing long/wide	0.01	0.10	0.02	122
Tug	0.17	0.12	0.14	1885
accuracy			0.68	69676
macro avg	0.25	0.27	0.25	69676
weighted avg	0.71	0.68	0.69	69676

LightGBM has an accuracy of 68% only and is not performing as well as random forest.

6.4 LightGBM classifier

	precision	recall	f1-score	support
Cargo	0.67	0.94	0.78	37524
Dredging	0.28	0.01	0.02	1062
Fishing	0.74	0.73	0.73	4986
HSC	0.76	0.74	0.75	709
Law enforcement	0.00	0.00	0.00	320
Military	0.17	0.00	0.00	1416
Passenger	0.63	0.18	0.27	3298
Pilot	0.67	0.73	0.70	789
Pleasure	0.51	0.05	0.09	625
Port tender	0.00	0.00	0.00	57
Reserved	0.00	0.00	0.00	144
SAR	0.72	0.60	0.66	709
Sailing	0.00	0.00	0.00	370
Tanker	0.68	0.31	0.43	15452
Towing	0.50	0.02	0.04	208
Towing long/wide	0.00	0.00	0.00	122
Tug	0.67	0.73	0.70	1885
accuracy			0.68	69676
macro avg	0.41	0.30	0.30	69676
weighted avg	0.65	0.68	0.62	69676

Linear SVM also has an accuracy of 68% similar to LightGBM and is not performing as well as random forest.

6.5 Linear SVM Classifier

In [33]:
```
rf = RandomForestClassifier(random_state=7)

# specify the hyperparameters and their grid values
# 4 x 3 x 2 = 24 combinations in the grid

param_grid = {
    'n_estimators': [50, 100, 200, 400],
    'max_depth': [6, 10, None],
    'min_samples_split': [2, 5]
}

# using 5-fold cross-validation
grid_search = GridSearchCV(rf, param_grid, cv=5, scoring='f1_macro', return_train_score=True, n_jobs=-1)

start = time.time()
grid_search.fit(X_train, y_train)
end = time.time() - start
print(f"Took {end} seconds")
```

```
Took 2315.9793124198914 seconds
```

In [34]:
```
print('best parameters from grid search:')
print(grid_search.best_estimator_)

print("best f1_macro score:", grid_search.best_score_)
```

```
best parameters from grid search:
RandomForestClassifier(min_samples_split=5, n_estimators=200, random_state=7)
best f1_macro score: 0.9100639583811295
```

Best hyperparameter values are:

6.6 Fine tuning random forest classifier using grid searchcv

7.6.2 training the RF classifier using best hyperparameters

In [36]:
```
%%time

# random forest classifier object
rf = RandomForestClassifier(n_estimators=200,\
                            min_samples_split=5,
                            max_depth=None,
                            n_jobs=-1
                           )
rf.fit(X_train, y_train)

# Prediction on test
rf_pred = rf.predict(X_test)

# performance of random forest classifier on test
print(classification_report(y_test, rf_pred))
```

6.7 Training random forest classifier with best hyperparameters