# Predicting Port Tugboat Operations for Arriving and Departing Vessels Using Machine Learning

Master's Thesis in Technology
University of Turku
Department of Computing
Software Engineering
2022
Adrian Borzyszkowski

Supervisors:
Petra Virjonen
Jukka Heikkonen

UNIVERSITY OF TURKU
Department of Computing

Adrian Borzyszkowski: Predicting Port Tugboat Operations for Arriving and Departing Vessels Using Machine Learning

Master's Thesis in Technology, 74 p., 4 app. p.
Software Engineering
February 2022

Today, predicting the number of tugs required to assist in a towing operation many days in advance is difficult. Towing operations, being a complicated process, are prone to human errors and conflicts, which can have severe financial consequences for all parties involved.

In this thesis, a method for extracting port tugboat operations for incoming and outgoing vessels is proposed. Using the obtained tugboat operations dataset, a machine learning model is built in order to predict the number of tugs required to assist in a towing operation. The data used is a year of historical Baltic Sea AIS data and weather data from nearby weather stations near the two analysis ports.

The recommended ideas and their implementation were a success from a performance standpoint. The proposed method for extracting towing operations detected the vast majority of towing operations within the analysis area. The obtained tugboat operations dataset was then used during the model construction phase. The obtained models are port-specific. One of the models achieved an overall accuracy of 87.0%, while the other achieved an accuracy of 91.5%.

The results demonstrated that it is possible to develop a viable predictive tool for tugboat operations. When deployed, the proposed method will enable port and tugboat operators to make faster and more efficient decisions, resulting in increased operational efficiency in the port area.

Keywords: tugboat, port operations, data processing, AIS, machine learning, XG-Boost, K-nearest Neighbors

# Contents

**Appendices**

# List of Figures

# List of Tables

# 1 Introduction

The topic of this thesis stems from problems and challenges in the area of marine logistics. Ports now play a significant role in the worldwide transportation of commodities, people, and products. It is common nowadays to see businesses upgrade their processes in order to improve resource and time management. However, the maritime logistics industry is considerably behind in this respect and is only now beginning to catch up by adopting contemporary digital marine system infrastructures, including Artificial Intelligence (AI) driven solutions [1], [2]. The marine industry is quickly evolving due to the rapid pace of technological change, such as digitalization and automation, and for this reason, may become more efficient. As a result, there might be significant financial benefits to integrating digital technology into the whole maritime system. Nations like Spain, South Korea, and many more plan to upgrade their seaport infrastructures in order to improve efficiency, safety, and sustainability at sea, ports, and on land [3], [4]. According to the study on the digitalization of Finnish ports, one of the most critical areas of digitalization in ports is improving the open data, which refers to freely available information in machine-readable format [5]. Open data enables application developers to create a range of port-related services. Such services have the potential to significantly improve port traffic flow and performance. They aid in decision-making by offering data-driven information in addition to empirical knowledge. A good example is applications based on Automatic Identification System (AIS) data from ships, which are already important tools in ports. AIS data-driven applications support situational awareness and decision-making.

One of the procedures in the domain of port infrastructure that is yet to be updated and enhanced is the towing operations for arriving and departing vessels. Ship towage operations involve mooring and unmooring of vessels at a port and are performed using port tugboats. Today, determining the number of tugboats necessary to assist the vessel is complicated and requires taking into consideration a range of circumstances. Additionally, the choice on the number of tugboats to assist the vessel is often subjective and is dependent on a port pilot a vessel captain. Human factor analysis in marine transportation [6] indicates that the human factor is a major element in maritime accidents. It is also important to underline that towing operations are often integrated into pilotage operations. Pilotage services provide guidance for mooring and unmooring vessels and ensure safe and efficient passage of vessels near the port areas and canals. During pilotage, a ship's captain, in consultation with a port pilot, manages the whole operation and determines which tugboat services will be used. In pilotage operations, human error is a possibility [7]. Inadequate communication between the bridge and pilots, a frequent source of incidents during pilotage operations, may have significant consequences. The implications of such accidents are concerning, because incorrectly executed pilotage and towing operations may have financial and even safety consequences. In pilotage and towage operations, human errors are influenced by the safety environment on the command bridge. Personality conflicts, pressure from ship owners, and lack of trust are all examples of factors contributing to the hostile work environment. Towing and pilotage operations will be explained in detail in the following Chapter 2.

AI systems are far more capable of doing a wide variety of repetitive activities than people are. Such systems always handle data in the manner in which they are configured. This frees up time for workers' to focus on strategic decisions directly connected to the core operation of the organization. The better the data quality is, the more likely employees make smarter decisions that save organization money. By introducing corporate procedures with AI, it is also possible to discover and eliminate inefficient activities. AI tools extract important data that would be impossible to get manually.

Machine learning is a subfield of AI. It is the study of computer algorithms which learn the rules for solving a given problem using the collected data and their own mistakes. A machine learning-based prediction tool for port tugboat operations may assist port decision-makers and port tugboat providers in optimizing and enhancing the marine logistics industry. Predictive tools introduction may result in more efficient port call planning, successful marine operations in port areas, and better fleet utilization. Additionally, it could also help reduce emissions associated with port operations and possibly improve safety.

This thesis will explore a solution utilizing machine learning-based techniques for improving the decision-making process associated with tugboat operations in order to enhance it. The goal is to build a machine learning model that can predict the number of tugboats required to assist an arriving or departing vessel in port. This is not a trivial problem and requires an assessment of many factors, including the present weather conditions on the day of the event, the vessel's specific characteristics, vessels positions at sea, unique port characteristics, the availability of tugs, and many more. This issue is mathematically intractable and that is why it requires the development of a machine learning-driven solution.

While enhancing tugboat operations is just one step of many into a long process of upgrading and modernizing logistics operations, the end result has the potential to significantly benefit the maritime logistics industry in many ways. The thesis subject was suggested by Awake.AI [8], a Finnish firm that specializes in assisting maritime operators implementing contemporary system infrastructure. Their customer base is diversified, and they collaborate with a variety of international port operators, authorities, and logistics organizations. Awake.AI has developed a platform for maritime operations that enables operators to make faster and more informed decisions with enhanced situational awareness. Port authorities and operators can use Awake.AI solutions to optimize all port operations using machine learning to accurately estimate a vessel's ETA and ETD (Estimated Time of Arrival and Departure); at the same time, ship operators and cargo owners can take use of situational awareness. Port actions are more secure and efficient in Awake.AI's perspective. They progress toward automation of operations, cranes, and ships. Their objective is to help the transition to sustainable and intelligent maritime logistics in collaboration with their partners, with the goal of reducing global $CO_2$ emissions by 10% of due to shipping by 2030.

## 1.1 Literature Overview

International marine transportation has grown tremendously in recent decades. However, there have been few research on how machine learning may be used in this industry. To my knowledge, there are no studies on towing operations prediction using machine learning, thus now is a good moment to conduct academic research on this topic. This section briefly discusses some of the previous researches on this topic that are conceptually similar or related to this thesis.

An important and often addressed problem in academia is the topic of predicting ETA and next destination of a vessel using machine learning methods. Recent methods, such as this 2020 research, incorporate AIS data-driven vessel destination prediction utilizing a random forest approach [9]. Although this article does not directly address prediction of tugboat operations, it provides an excellent overview of the process of AIS data usage in machine learning-driven applications in the marine logistics field. The paper offers an AIS data-driven solution for general vessel destination prediction in global shipping services. In this paper, a random forest model is built using AIS data.

The problem of ETA and destination prediction was approached using Neural Network and Graph Clustering methods. The research team was able to achieve an accuracy of 97% for port destination classification and 90% for ETA in minutes for the DEBS Grand Challenge 2018. Their final prediction model consists of an ensemble of machine learning models and a deep learning model [10]. In the thesis made by Orgaz Expósito, the destination prediction model for European waters achieved 90% accuracy, while the model for ETA prediction achieved a low mean error in relation to the voyage duration [11].

Another useful and inspiring recent paper is an aid decision-making model for ship detention based on Extreme Gradient Boosting (XGBoost) [12]. Paper proposes a tool that is used to aid flag state control officers (FSCOs) in accurately determining whether an inspected ship is "detained". The issue at hand is a binary classification problem, while the problem at hand in this research is a multi-class classification problem, which are fundamentally comparable; multi-class classification will be explored more in Chapter 4 on machine learning. The paper makes use of a mix of XGBoost and other predictive algorithms. XGBoost will also be utilized as a predictive model in this research. The paper enabled the development of the intuition for XGBoost and the assessment techniques for this prediction approach.

Meng et al. used AIS data to analyze vessel traffic characteristics in the Singapore Strait [13]. The research team discusses the benefits and drawbacks of AIS. AIS is simple to access and contains significantly more information than traditional Vessel Traffic Service (VTS) data. The disadvantage of AIS is that data may occasionally be missing. This could be a result of a hardware inaccuracy or simply a human error. Additionally, the study proposes various methods for improving the quality of AIS data, such as minimizing data entry errors and offers additional sources to refine AIS data. Numerous other applications make use of the benefits of AIS data. There is an increasing amount of research that makes use of its advantages. Yang et al. provide an exhaustive review of studies relating to the use of AIS [14]. AIS is used in a variety of analyses, including ship behavior analysis, environmental assessment, trade analysis, ship and port performance, studies on arctic shipping, and many more. AIS not only increases the visibility of shipping activities but also makes them more analyzable.

There have been relatively few studies on tugboat operations. Paulauskas and Paulauskas discussed the mooring and unmooring operations for tugboats in seaports [15] from the theoretical point of view. The study provides an in-depth academic overview on how mooring and unmooring operations are carried out and what forces are involved.

Tugboat operations were also the topic of an insightful analysis published in May 2020 [16] by Chen et al. Study aimed to identify the maneuver services and analyze the characteristics of tugboat activities in the port area using AIS data. One of the study objectives was to extract the tugboats performing manoeuvre services based on vessel kinematic features reflected in AIS data. The paper was primarily used to provide an academic perspective on port tugboat operations, particularly berthing and unearthing maneuvers. One of the major findings of this study was that vessel length has a significant effect on the tugboat fleet's utilization. It is important to emphasize, however, that none of the elements discussed in the paper were used in this study and served only to complete the academic picture of towing operations.

The existing studies on tugboats include also the problem of tugboat scheduling. Zhen et al. proposed an approach for tugboat scheduling in hinterland barge transport [17]. The difference between river and harbour tugboat operations is that instead of maneuvering a large ship, tugboats on the river usually tow multiple barges at the same time.

## 1.2 Research Motivation

A tugboat predictive tool may be adapted to assist different stakeholders and operational sectors. Tugboat prediction tool may assist port decision-makers in the marine logistics industry in obtaining the most precise prediction for optimizing logistic events. Additionally, such predictive technologies may aid in more efficient and accurate port call planning, marine operations in port areas, identifying and optimizing fleet utilization, lowering emissions associated with port operations, and possibly improving safety in hazardous circumstances.

Port authorities may increase the port throughput by making faster and more accurate decisions and also decrease operating costs and emissions by optimizing the use of tugboats in the port area, which is particularly advantageous in busier ports where resource allocation must be carefully planned in advance. Additionally, a tugboat service providers could also benefit from such application by allowing to allocate resources in places where needed. This would be favourable especially in busier ports where port-calls are planned many days in advance.

Utilizing a tugboat predictive tool may possibly enhance overall ship safety in port by providing additional information on the number of tugboats, which is especially useful in case the port crew is inexperienced.

## 1.3 Research Aims

Two research questions have been established in order to solve the difficulties described above in relation to the evaluation of vessel towage assistance using a machine learning methodology. In this thesis, the aim is to answer the following study questions:

RQ 1. Is it possible to extract port tugboat operation events for incoming and outgoing vessels using AIS data with sufficient accuracy?

RQ 2. Is it possible to develop a machine learning model for estimating the number of tugs to assist in a towing event?

The first objective of the thesis is to create a dataset of tugboat operations using AIS and weather data. This will be accomplished by extracting tugboat events from AIS data and then assigning each towing operation event and its characteristics including current weather conditions to a single data-point. A tugboat operation event is defined as one that begins when the vessel enters the port area or begins unmooring from the pier and ends when the vessel exits the port area or moors to the pier. Typically, a vessel will have two tugboat operations: when it enters the port and when it exits the port. The time intervals between these events can vary significantly. Once a dataset has been constructed and tested, the second objective will be to predict the number of tugs required to assist incoming or departing vessels in port using the obtained dataset.

The number of tugs assisting in the event is critical to correctly identify, as it plays a significant role throughout all thesis phases. This is the parameter that teaches the model on how to predict. Once correctly identified, adequate data will be obtained, enabling the model to be adequate as well. A good model requires the incorporation of sufficient related features to differentiate the various prediction outcome schemes. For example, a simple towing operation event requiring only one tug may have entirely different characteristics than a more complex event involving two or more tugs.

## 1.4 Thesis Structure

This research will focus on two Finnish seaports with sufficient maritime traffic to allow for the collection of data. After conducting an initial study of Finnish ports, it was found that the ports of Rauma and Vuosaari are viable candidates. Two ports were chosen instead of one to allow for a more detailed study of the results; each port is unique, and any potential differences or similarities of the prediction results should be highlighted and discussed. Data processing and analysis are done using Python, e.q. jupyter-notebook. It is important to emphasize that the objective of this research is not to maximize the performance of the models but to determine if research questions can be answered and whether the concept for such a system works, and that any possible enhancements of the processing and predictive system are going to be part of subsequent studies or projects.

To begin with, one must have a deeper understanding of marine logistics and seaport structure in general, as well as the many kinds of ports, their layouts, and the various types of ships and tugboats. Additionally, it is essential to understand how towing operations in the port area are conducted and the specifics connected with them, what the distinctions and goals of such maneuvers are, and what is necessary while conducting such actions. Moreover, it is critical to grasp the rationale behind the choice about the number of tugs, how it is made, and what are the major variables influencing such a decision. Naturally, understanding maritime logistics requires an understanding of how AIS operates. AIS and wind data were utilized to build the dataset for the learners during the data preparation phase. Understanding how AIS operates will aid in the discovery of valuable characteristics that will be utilized to construct the dataset. All of this enables a more interpretable research problem and the capacity to define the project's scope precisely. This stage is critical since any significant outcomes may result in effective methods and experiments throughout the project's implementation phase. In this instance, the hypothesis is that this problem can actually be addressed using machine learning techniques.

Predictions of tugboat operations are made using K-Nearest Neighbour and XGBoost techniques. Initial iterative analysis was performed using the K-Nearest Neighbor method to enhance the quality of the data processing since XGBoost training time is much longer on average. Once an acceptable level of data quality was obtained, the final prediction was carried out using the XGBoost algorithm. K-Nearest Neighbor results served as a baseline for improvements.

The structure of the thesis is the following: Chapter 2 gives essential theoretical background on AIS and tugboat operations in general, as well as describe the method of selecting the number of tugs for a towing event and the rationale for picking the ports of Rauma and Vuosaari. Chapter 3 is primarily concerned with the processes of acquiring and assessing the quality of the tugboat events dataset. Chapter 4 introduces fundamental machine learning concepts that have been utilized in this study. Chapter 5 presents the obtained results. Chapter 6 analyzes and discusses the obtained results and gives suggestions for possible future improvements. Finally, Chapter 7 concludes the thesis.

# 2 Background

This chapter will cover the background information necessary for comprehending the topic assessed in the introduction as well as detail the scope of the analysis. First, the fundamentals of the AIS will be discussed. Necessary information will be given regarding tugs and towing events in general. It is also important to understand the reasoning behind the decision on the number of tugs to be utilized in a towing operation. Finally, the scope of the analysis will be established, which will include the rationale for selecting the ports as well as the specifics of the selected ports themselves.

## 2.1 Automatic Identification System

AIS is an automatic communication system developed at the end of the 20th century and widespread at the beginning of the 21st century [18]. AIS provides automatic data exchange, useful for avoiding collisions between vessels and ship identification for coastal systems monitoring ship traffic. AIS was created primarily as a tool to prevent collisions at sea. Currently, almost all ships are equipped with an AIS transponder in accordance with Convention on Safety at Sea (SOLAS) [19, Chapter V, Regulation 19]. AIS allows vessels and coastal stations equipped with similar transponders to transmit their own position and traffic parameters and to receive such information from other units.

The main task of the system is to improve the safety and security of navigation by enabling the monitoring of traffic in the water area. Any vessel, ground station, or satellite within range of AIS transponder can receive the messages. Thanks to this, it is possible to monitor the movement of ships wherever they are. Ground stations detect AIS signals from vessels near shore, while satellite-detected AIS enables monitoring of vessels far offshore. Because the AIS signal is broadcast over an open frequency, other vessels, coastal states, and private businesses have access to the data collected from ships.

All ships over 300 GT engaged in international trips, as well as passenger ships of any size, are required to be equipped with the AIS system mandated in SOLAS. Gross tonnage (GT) is a measure of a vessel's overall internal volume and it is used to determine things such as safety rules, registration fees, and port dues [20]. Tugboats must also have a transceiver installed. The EU law expanded the SOLAS regulation's material scope by requiring fishing vessels, for example, to be equipped with an AIS system [21].

Continuously updated data from AIS may be used to determine a vessel's position at any given time, and a well executed analysis of this data should be able to determine the ports visited by a given ship at any given time, as well as the path taken by those ships.

AIS transceiver broadcast the following types of data messages: static along with voyage variables presented in Table A.2 and dynamic variables in Table A.3. Ship MMSI number (Maritime Mobile Service Identity), IMO number (International Maritime Organization), reference numbers, ship type, and vessel name are all examples of static information. Dynamic information, on the other hand, includes the ship's location and positional time.

It is important to understand the concept of AIS and how it works because data obtained for this analysis will mainly be based on the AIS data. The quality and the human involvement on the AIS has been addressed in numerous studies [22], [23].

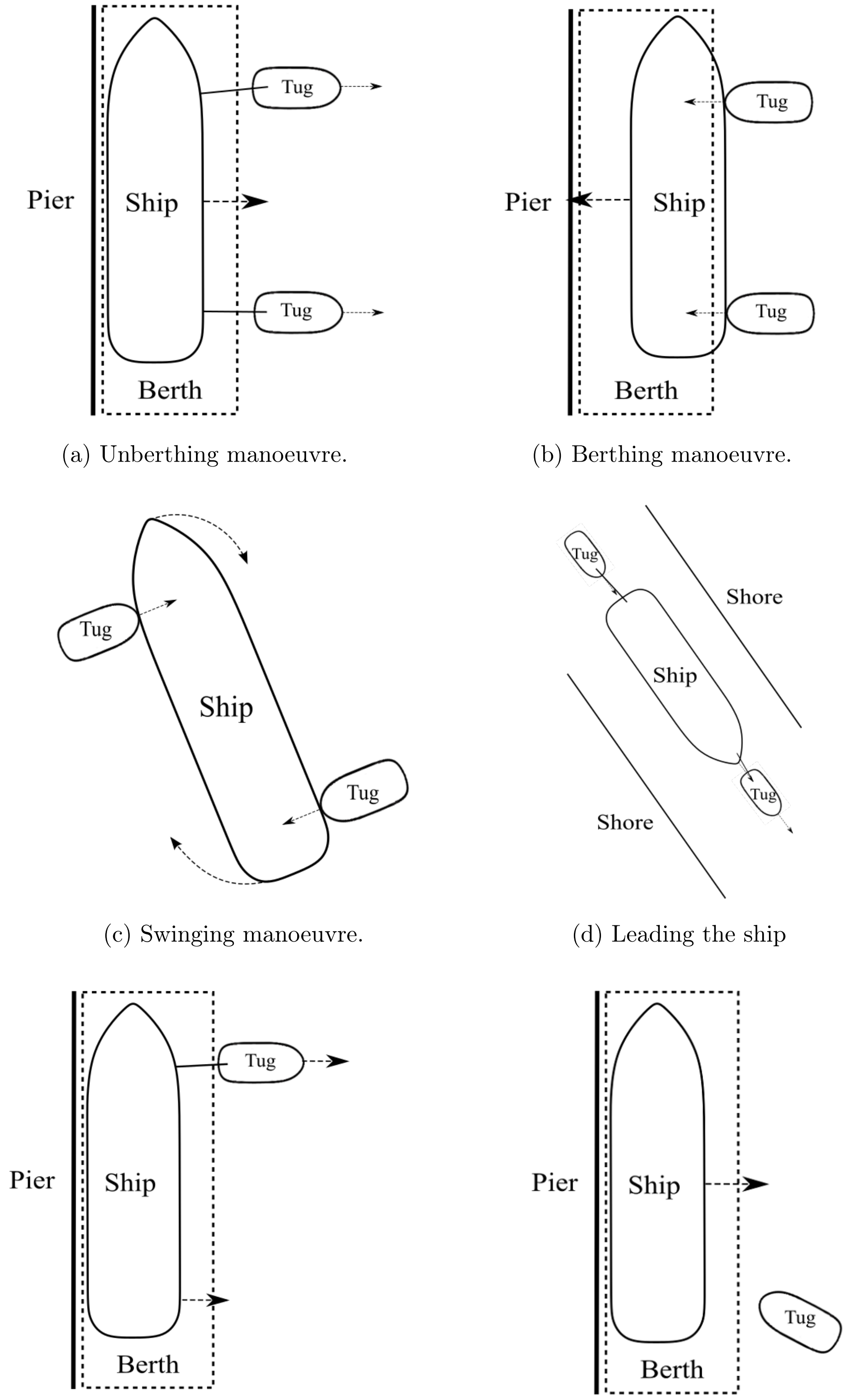
Because the AIS system lacks built-in validation mechanisms, it is up to a coastal officer to review and validate the accuracy of the transmitted data on a regular basis. The quality and reliability of the gathered AIS data will be further addressed in the following Chapter 3.

## 2.2 Tugboat and towing operations

A tugboat (or tug) is a very strong and mobile work vessel used in water transport, specially built for towing operations. Tugs with engines that have disproportionately high power and traction in relation to the size of the vessel are capable of maneuvering ships that are many times heavier than themselves. The size and power of the tug can vary greatly, and this is mostly determined by the tasks for which it was designed and built. Towing may result from various situations, both planned and unplanned. Towing a specific vessel from one place to another, resulting from an agreement between the owner of the vessel and the tugboat, is the most common task of a tugboat.

Tugboats are multi-functional vessels that, while designed for towing, can also be utilized for other purposes, such as unanticipated rescue, transporting, salvage, ice-breaking, or dredging. However, this study will focus only on tugboat assistance in the port areas, including one or more tug maneuvers and escorting.

There can be differentiated numerous techniques for carrying out towing operations that depend on the location where the action was carried out, as well as the wind forecast for the duration of the action. Figure 2.1 depicts various tugboat assistance methods in the port area. Figures *a* and *b* represent unmooring and mooring manoeuvres that require two or more tugs. Tugs pull or push the ship with sufficient force to move it in the desired direction. To keep the ship parallel to the pier, both the forward and after tugs must be constantly controlled. In these cases, tugs are used to avoid making hard contact with the pier. Figure *c* represents a



(a) Unberthing manoeuvre.

(b) Berthing manoeuvre.

(c) Swinging manoeuvre.

(d) Leading the ship

(e) Single tugboat assist the ship equipped with thrusters.

(f) Escorting tugboat ready to assist the ship if needed.

Figure 2.1: Different methods of tugboat assist in the port area.

swinging manoeuvre, where two or more tugs are used to rotate the ship. When a ship is unable to rotate safely on its own near a port pier or quay due to a lack of thrusters, a swinging maneuver is needed. Tugs push in opposite directions in their respective places. Figure *d* illustrates a situation in which a ship arriving or departing from a busy river port may require tug assistance to turn around tight bends by controlling the ship's bow and stern. For ships equipped with thrusters, for example, it will be unsuitable to use two or more tugs. In such situations, the ship can either maneuver without tugs or utilize only one tug if necessary (figures *f* and *e*).

## 2.3 Choosing the number of tugs

A ship captain will receive advice and assistance from a port pilot upon entering the port area. The pilot will board the ship while entering or leaving a port. The pilot acts as an advisor to the captain and as an expert on the local waters and their navigation. The port pilot will be aware of the limits of the tugs that will be employed, the competence of the tug crew, the swinging room that will be available, and the needed angle and distance of the towing line in order to perform the manoeuvre as efficiently as possible. The purpose of pilotage is to ensure the safe entry or exit of the ship. The Captain and pilot cooperate with the tug teams supporting mooring or unmooring ships. The pilot will advise the captain on the number of tugs needed for the operation, but the captain and the pilot must agree on the final decision. If the captain and the pilot cannot agree on the number of tugs needed for the operation, the port pilot has the option to refuse pilotage.

It is also critical to underline that any hesitation on the part of the ship captain should be considered as a reason to hire tugs. Personal ambition and the desire to prove one's skills and abilities or the costs of hiring towing assistance should not be a reason for risky decisions and refusal to use tugboats. Towing operations can be expensive, and it may be in the captain's or shipowner's best interests to avoid them, but one must keep in mind that the cost of towing operations is negligible in comparison to the price of a potential marine accident. Port tugs play an important role in port navigational safety. Many ships now have thrusters, which can take over some of the tug responsibilities, but many ships, especially large ships, do not, which is why tugs are so crucial for improving navigational safety [24].

The use and the number of tugs is determined using following considerations:

- Regulations of the port stating the limitations of the maneuvering area for a certain types of vessels. The Port Authority may order mooring or unmooring vessels to use towing assistance to avoid damaging the quay or any cranes located on the quay.

- Wind conditions during the event. The direction and angle at which these forces will impinge on the ship's hull are critical. Excessive wind force and an unfavorable wind direction could drive the ship in an undesirable direction.

- Type of propulsion system equipped onto ship. When a vessel's maneuvering speed is too high, it loses course stability at low speeds.

- State of the ship's propulsion systems. Malfunction of the propulsion system, steering, or inability to use the anchors could greatly affect the maneuverability of the ship.

- Upon the request of the captain, who, in consultation with the port pilot, determined that the maneuver was both too difficult and too risky under the current circumstances and conditions.

- Availability of the specific tugs. Specific tugboats may not be suitable for towing certain types of ships.

- The amount of cargo onboard the ship.

## 2.4 Port Selection

As mentioned in Section 2.2 the analysis scope is the tugboat operations in the port areas, including one or more tug maneuvers and escorting. This analysis will focus only on events of incoming or outgoing ships in restricted areas, such as ports and channels. Initial analysis was done for a certain list of Finnish ports. The project involves the extraction and analysis of towing events from the ports of Rauma and Vuosaari. The ports of Rauma and Vuosaari were chosen for this study for a number of reasons:

- The objective of this study is to establish a tugboat operation prediction for a set of Finnish ports.

- According to initial traffic analysis of a group of Finnish ports, the ports of Rauma and Vuosaari are among the busiest, meaning that the number of potential towing operations will be higher, perhaps enhancing data quality.

- According to the 2017 Finnish Transport Infrastructure Agency Finland has 91 active tugboats in use [25]. Initial analysis revealed that the port of Vuosaari used a total of 19 tugboats, seven of which were extensively utilized. In the case of port of Rauma, a total of 21 tugs were at some point used, of which eight were used more than others. It is noteworthy, that the tugboat service provider could change the tug operation location to other ports. It is quite common for tugs to perform tasks in other ports before returning to their home port. As a result, there is an indication that tug services were likely used frequently in these ports.

- Another motivation for choosing Rauma and Vuosaari was the opportunity to evaluate the final quality of the data using the ground truth data provided by Alfons Håkans [26] dataset. Alfons Håkans is the main tugboat service company operating i.a. in the ports of Rauma and Vuosaari. The quality assessment will be discussed in the Chapter 3.3 concerning the data quality evaluation.

### 2.4.1 Port of Rauma

The Port of Rauma (LOCODE: FIRAU) is a cargo port located on the southern east side of the Gulf of Bothnia. LOCODE is a geocoding system that is managed by the Economic Commission for Europe of the United Nations (UNECE) and it is assigned to locations critical to commerce and transportation, such as ports [27]. The Port of Rauma is the largest container port on the west coast of Finland, accounting for more than 70% of containers traveling through the west coast ports between Turku and Tornio. In 2019, the port handled 1081 vessels in total, managing around 5.8 million tons of international cargo. The main export materials were: paper and cardboard, pulp, sawn goods, and the main import materials were round wood and kaolin. Port of Rauma has seven different quays. The most often used quays for container ships are the New Container quay, which is 350 m in length and 13.6 m in depth, and the Iso-Hakuni quay, which features five loading berths and a depth of 11.0 m [28]–[30]. Figure 2.2 gives an indication of the port arrangement.

The port of Rauma was designed to handle vessels up to 255 m in length, 32.2 m in breadth, and 12 m in depth. The vessel's maximum permitted speed in close proximity to the port is ten knots. Whenever the ship's length exceeds 210 m, the maximum allowed speed of the wind gusts during the mooring or unmooring is 15 m/s during the day and 12 m/s at night. Whereas for ships with length under 210 m wind gust limit is more loose with 18 m/s during the day and 15 m/s at night. Pilotage and therefore towage is discontinued when the wind speed exceeds 20 m/s [29].

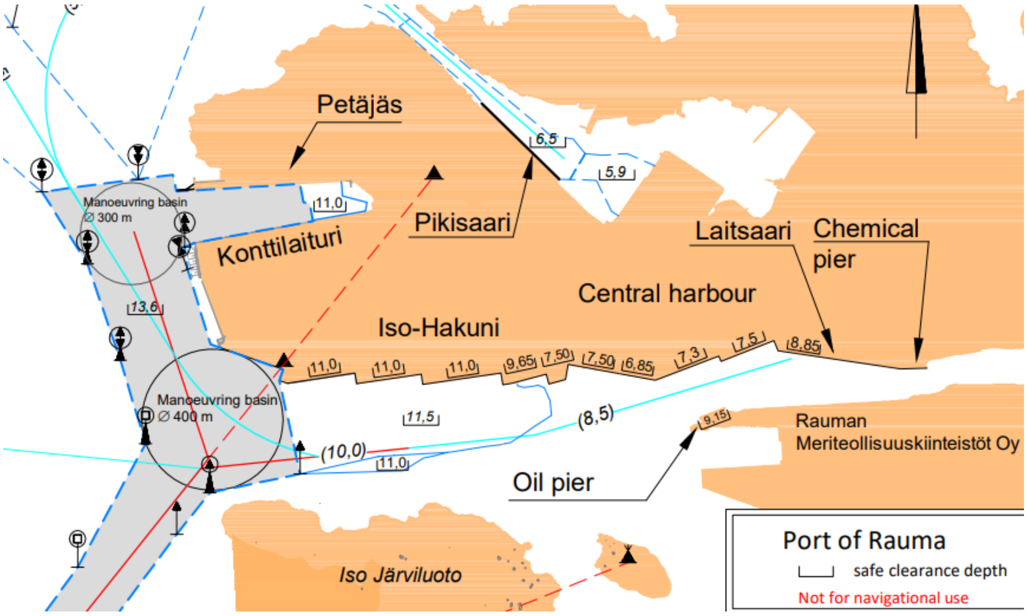


Figure 2.2: Port of Rauma. Port of Rauma has seven quays with each having specific clearence depth. Ship arriving from southwest are allowed to perform swinging maneuver in region represented by the circle [29].

### 2.4.2 Port of Vuosaari

The Port of Vuosaari (LOCODE: FIVUO) is a cargo port located on the northern shore of the Gulf of Finland, approximately 15 kilometers east of Helsinki. The port is a major port for container and trailer traffic. In 2019, the port handled 2 354 vessels in total, managing around 14 million tons of international cargo. Vuosaari's traffic is fairly balanced. In 2018, exports accounted for 49.3% of the total transported weight, while imports accounted for 50.7%. [31], [32]. Figure 2.3 gives an indication of a port arrangement. It is also worth noting that vessels with the destination of FIHEL sometimes enter the Vuosaari port instead of port of Helsinki.

Port of Vuosaari by having seven different quays with various piers and extensions is able to handle much more traffic than the port of Rauma. The port of Vuosaari was designed to handle vessels up to 230 m in length, 33 m in breadth, and 11 m in depth. Similarly to the port of Rauma, the vessel's maximum permitted speed in close proximity to the port is ten knots. The Longest quay E is 749 m long with a clearance depth of 12.5 m. The maximum allowed wind speed for vessels over 200 m in length is 15 m/s [33]. Typical wind directions and strengths for both ports will be discussed in Chapter 3.1.2.

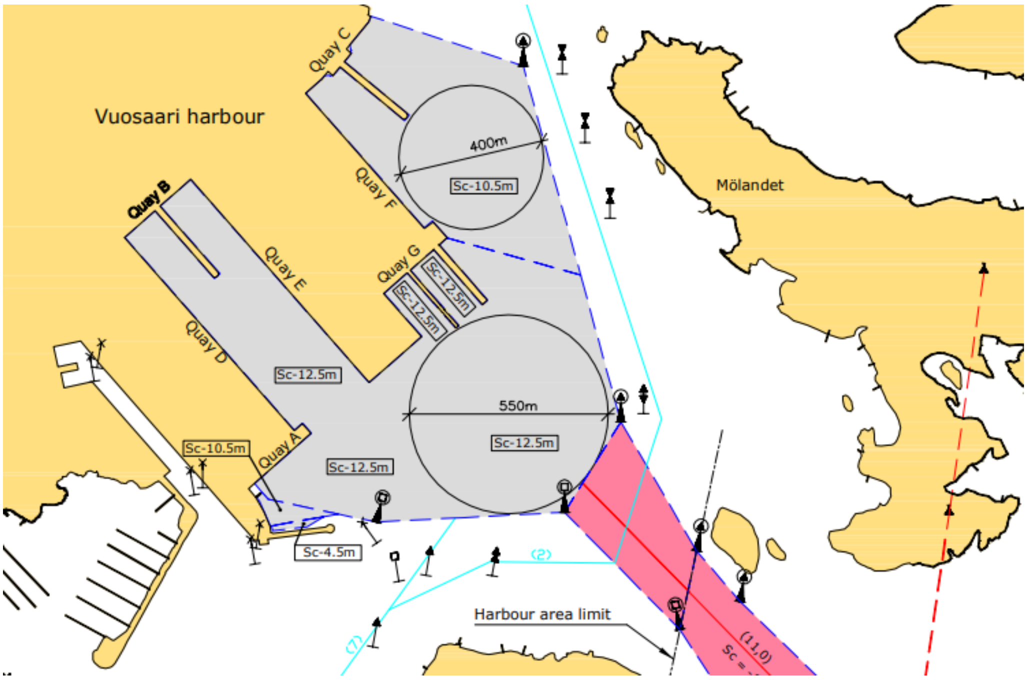


Figure 2.3: Port of Vuosaari. Ship arriving from southeast are allowed to perform swinging maneuver in region represented by the circle [33].

.

# 3 Data preparation

In order to achieve goals set in the Chapter 1.3 a specific set of variables out of the gathered AIS and weather data were selected. Selected AIS and weather features could be beneficial in picturing the occurred event and with their help it could be possible to determine whether ship needed tug assistance or not. This will allow to build a dataset for the machine learning classifiers. Selected AIS features which mainly consist of positional information and dimension of the vessel help in visualizing the path that has been taken. Vessel paths will enable interaction analysis between the ship and the vessel, determining the maximum number of tugs participating in the towing operation. Wind features describe the conditions during the event and include the direction of the wind and force. While the AIS and weather datasets enabled the use and extraction of features that appeared beneficial during the planning phase of the project, new ideas for additional features were considered throughout the study's development, which could potentially improve the predictive performance of the model. This topic will be discussed in detail in Chapter 6.3.

Only AIS and weather data were available during the data preparation phase of this research. While the features utilized and extracted for the needs of this thesis reflect the most critical considerations identified in Section 2.3, not all considerations for the tug number selection can be translated out of the used data and need the use of alternative sources of information. For example, it is impossible to determine the number of personnel and tugs available at a specific location using AIS and would require the use of alternative sources of information.

In this research no feature selection was employed since one of the goals was to determine which features are advantageous and which are not. Any subsequent research may include feature selection in order to increase prediction performance. Chapter 4.11 will describe in depth how to evaluate the relevance of features.

This chapter will focus on describing the data that has been used in this analysis. Additionally, the processing of the data will be explained. Lastly, the quality of the obtained final dataset will be assessed. In the following analysis *"ship"* refers to a tanker, cargo or passenger ships, whereas *"vessel"* refers to either a ship or a tugboat.

## 3.1 Data used

The importance of wind conditions in decision-making has been stressed in the Section 2.2 on towing events, and as a result, towing events must be examined from the perspective of wind conditions as well. The gathered AIS and wind data starts from July 1, 2019 up to June 30, 2020.

Data for this project was built mainly from two sources: the source for AIS messages was Digitraffic.fi [34], and the source for weather data used in this analysis originated from the Finnish Meteorological Institute (FMI) [35]. Both the AIS and weather data used for this analysis were provided by Awake.AI.

### 3.1.1 AIS data

The gathered AIS data is a collection of AIS messages starting from 01.07.2019 to 01.07.2020. Original AIS data has been narrowed down to positional information about vessels in the Baltic Sea. Due to the sheer size of the data, it has been divided by month, with each containing over 20 million messages. Division of the data into smaller portions speeds up and simplifies the processing of the data by enabling parallel execution of the data-processing pipeline. The table below describes the AIS features used in the analysis:

Table 3.1: AIS features used in the analysis

| Information | Description | Unit | Example |
|---|---|---|---|
| MMSI | Unique ID of a vessel | integer | 249598000 |
| Ship type | Type of ship and cargo type | integer | 80 |
| Time stamp | Current positioning time of a vessel | UTC milliseconds | 2020-06-01 12:11:14.754 |
| SOG | Speed of vessel over ground | knots | 12.2 |
| Location | Current position in latitude and longitude coordinates | degree | (59.14416, 21.931645) |
| True heading | True heading of vessel | degree | 183 |
| Reference points | Reference points A, B, C and D are used to calculate the dimensions of the ship. A+B=length of the vessel and C+D=breadth of the vessel | meters | 104, 28, 16, 3 |
| Draught | Maximum present static draught | meters | 8.6 |

The type of vessel is recognised by the number of ship type in the AIS message. Messages with ship types 52, 31, and 32 are messages from tugboats. Messages with ship types from 60 to 69 are passenger ships. Messages with ship types from 70 to 79 are cargo ships. Lastly, messages with ship types from 80 to 89 are tankers. The type of the ship's cargo is presumed to be unimportant in this research, hence the ship types have been divided into classes 60, 70, and 80. The amount of a particular cargo type is directly correlated with the ship's present AIS message draught variable, though to what extent is unknown. However, it is safe to assume that if there is a correlation between the amount of cargo and the ship's current draught, the machine learning model may discover it. Draught is the vertical distance between the waterline and the bottom of the keel. The draught of the ship is a critical physical parameter that determines both the ship's stability and the maximum loads that may be carried without causing damage. The draught of a vessel increases as the weight of the vessel grows in order to support the additional load. Ships can adjust the draught by filling or emptying their ballast tanks. Ships behave differently depending on their draught; in general, ships with an exceptionally high draught have limited maneuverability, whereas same ships with a low draught may be more susceptible to environmental changes such as wind. Even relatively common natural phenomena such as waves or wind can pose a major danger to ship stability, and in extreme cases, may even result in the vessel collapsing. The draught of the ship may play a significant role in determining the number of tugs required for operation.

### 3.1.2 Wind Data

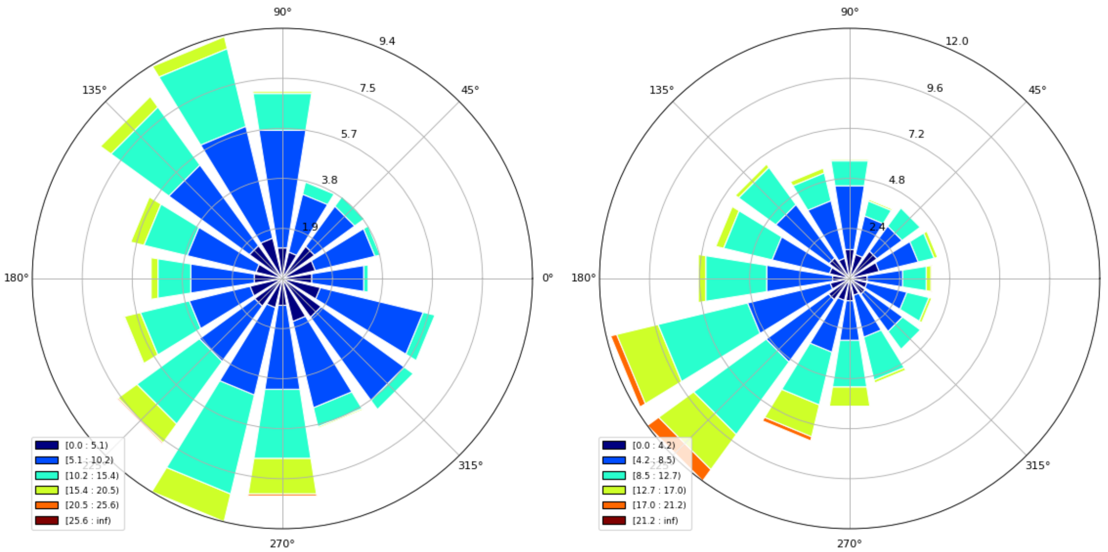
From the second chapter, it is known that wind conditions near the port are an important factor upon assessing the need for tugs for an upcoming ship event and also that ships require assistance from tugs, especially in windy conditions. FMI data is a collection of data produced by weather stations located near the shore. Wind data used in this analysis originates from two weather stations located closest to the ports of Rauma and Vuosaari. Vuosaari weather station is *Helsinki Harmaja* and Rauma weather station is *Rauma Kylmäpihlaja*. Similarly, as with AIS data, collection starts from 01.07.2019 and ends 01.07.2020. Due to the fact that FMI weather stations data included a break interval from 01.09.2019 to 09.09.2019, this period was excluded from analysis. FMI weather observations are updated hourly.

The following table details used wind information:

Table 3.2: Wind features used in the analysis

| Information | Description | Unit | Example |
|---|---|---|---|
| ws_10min | Wind speed, 10 minute average | m/s | 5.7 |
| wg_10min | Wind gust, 10 minute average | m/s | 6.5 |
| wd_10min | Wind direction, 10 minute average | degree | 107.0 |

Figure 3.1 represents the wind speed and direction for each studied port. In the case of port of Rauma, the wind generally blows from the southwest and northwest. In the case of the port of Vuosaari, the wind is primarily from the southwest.



(a) Port of Rauma wind data.

(b) Port of Vuosaari wind data.

Figure 3.1: Ports of Rauma and Vuosaari histogram with normed wind speed (displayed in percent) and direction. One year of data starting from 1.07.2019 to 1.07.2020.

## 3.2 Data processing

AIS data of each port is narrowed to a seven-kilometer radius from the center of the port area, meaning tug operations outside this area will not be detected. The seven-kilometer radius is a compromise between the area size and the tug operations pipeline's computation time. This area provides enough coverage of the port area, giving the necessary space for tug operation analysis, as tugs often meet ships at a meeting location 5 to 6 kilometers from the port area, and it reduces computation time by reducing the number of points to interpolate and analyze. The final model will predict tug operations only for ships that are coming or leaving the port. The scope of the project was to develop a machine learning model for incoming and outgoing vessels, and in-port operations were excluded from the analysis.

AIS messages may contain special or partly missing fields. In Maritime Anomalies and Safety research [36] it was discovered in an experiment that approximately 8% of the total ships operating in the Baltic Sea had missing data at some stages. Moreover, the study about AIS data reliability [23] has shown that especially AIS fields that are manually inputted are prone to missing values and need to be reviewed by the regulatory organizations. If any of the variables from Table 3.1 were missing, observation was not processed. Some features may also contain special values, i.e. SOG with 102.11 knots or true heading with 511 degrees. In the AIS, a common approach for dealing with missing data is to designate missing values with specific special values. Observations with variables having special values were regarded as if they were missing, and they also were left out of the analysis. With that said, the vast majority of the Digitraffic AIS data collected around the ports of Rauma and Vuosaari was normal, with no missing or special values.

The time intervals between two successive readings are represented in Table 3.3 below. Only moving vessels were considered; anchored vessels were excluded. The vast majority of readings in both Ports occur within a five-minute window, accounting for over 95% of total readings in both cases. However, anomalies are unavoidable. Even when anchored, vessels may transmit data as if they were slowly moving. Because ships are often anchored with a single anchor, the ship heading and transceiver position may shift gradually over time. When ships are anchored, ships sometimes leave circular data traces. This is why we can only estimate the overall percentage of time periods within a time range.

Table 3.3: AIS data time intervals between two consecutive readings.

| | % of total data points in that time range | |
|---|---|---|
| Time interval between two consecutive readings | Port of Rauma | Port of Vuosaari |
| [0, 5] minutes | 95.164 | 95.973 |
| [5, 10] minutes | 1.182 | 0.589 |
| [10, 15] minutes | 0.431 | 0.232 |
| [15, 30] minutes | 0.684 | 0.311 |
| [30, 60] minutes | 0.612 | 0.236 |
| > 60 minutes | 1.928 | 2.660 |

### 3.2.1 Interpolation

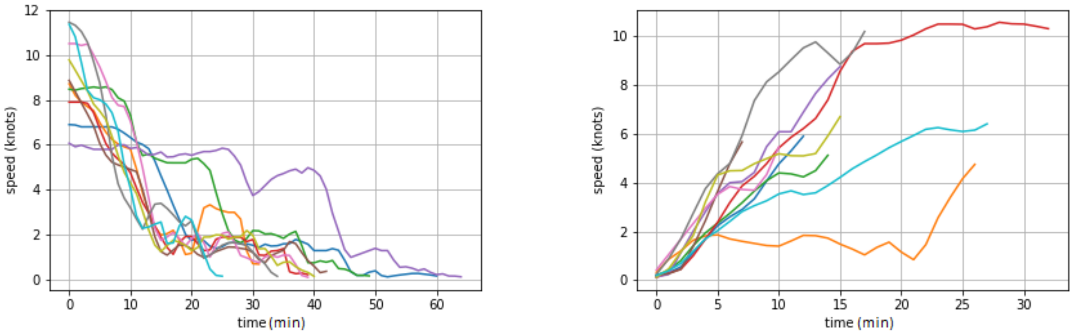
AIS data timestamp interval is infrequent, meaning that the time between two observations can change depending on the speed and maneuvers of the vessel. The frequency of transmission is determined by the ship's present speed and maneuverability [22], [37]. The Table A.1 shown in Appendix [22] provides the maximum duration between subsequent updates at different speeds, as well as the accuracy of that result. The maximum distance a vessel can travel between two updates is the accuracy guaranteed by the AIS. Naturally, AIS data has its own limitations.

Location data may contain some errors due to transmitter accuracy however, the difference is minor and in this study will be ignored.

To improve the accuracy of the calculated distance between the tug and the ship at a given time, observations must be first interpolated. This can be achieved by interpolating the path of the vessel by a constant time spacing. In this work time constant was chosen to be 60 seconds.

For each point to be interpolated, the closest neighbors in the original data set were searched and the interpolation was made using them. The location of the interpolated point was found by assuming that the ship travels at a constant speed and on the shortest route on the sphere between the two closest original points. The line between the two original points is a geodesic and can be calculated using the geographiclib WGS84 inverse function [38]. The ratio between interpolated time stamp and data-point pair time stamps is the same as the ratio between locations. Data-point pair is understood as a pair of locations *(latitude1, longitude1)* and *(latitude2, longitude2)*. The interpolated locations were obtained by matching the above-mentioned ratio on geodesic curve.

The difference between the real path and the interpolated path is affected by the assumption of the constant speed and the shortest route between the original points. If these two assumptions are not true, there will be an error that is proportional to the severity of the deviation from the assumptions. Let us assume, for example, the vessel travelling with 12 knots and the time spacing between interpolated points is 30 minutes. Within this time 30 minute time frame, the path of the vessel is not necessarily the shortest one and difference between the true and interpolated paths is significant, due to great speed and time spacing. However, the reduction of the time spacing and the speed will minimize the error between the true and the interpolated paths. Figure 3.2 below shows the speed of the incoming and outgoing ships during the towage. For reasons stated in Chapter 2.2 about towing operations, tug operations need to be precise and slow. Noticeable in Figure 3.2 is that the speed during the mooring decreases below six knots. The vessel will not have time to drastically change the location, due to low speed and time spacing.



(a) The speed of the incoming ships (10 ships)

(b) The speed of the outgoing ships (10 ships)

Figure 3.2: Speed of the incoming and outgoing ships during the towage. In case of incoming vessels the speed during the mooring drops below six knots. Whereas, the acceleration of the unmooring ship appears to be close to constant indicating that movement vector is not changing much.

Although the final interpolated path accurately estimates the movement of the vessel, a visual analysis of the vessel movement revealed several inconsistencies in the interpolated paths. For example, in some cases during close encounters, tug appeared to be within the ship. These potential inaccuracies in AIS locations, as well as in variations between real and interpolated trajectories, must be taken into account in the tug-ship interaction analysis.

Interpolation technique chosen to interpolate SOG was linear interpolation, which is a method that linearly generates data points between specified coordinate positions, assuming linear behavior between those coordinate positions. Interpolation of true heading is also linear interpolation, although it must be noted that, whenever one of the two true heading values is within the fourth quadrant and the second is within the first quadrant of a unit circle, the true heading within the first quadrant is increased by 360 degrees. Final interpolated true heading values are then obtained by applying the modulo 360 operation. For each interpolated data point, a wind observation with the closest time is selected.

### 3.2.2 Distance between the vessels

AIS data messages contain reference points, which describe the distance from the AIS transmitter to the four edges of the vessel. At first, the distance between the vessels was measured by the distance between their transmitters, but this method proved to be inaccurate. The true distance between vessels is shorter than the distance between transmitters because the location of the AIS transmitter is not the same on each vessel. Some vessels may have it at the front, some may have it at the back. The distance between vessels needed to be determined by the minimum distance between the edges of the two vessels, which gives a realistic representation of the distance.

The vessel's edges are estimated using a bounding box; the width and the height of the bounding box correspond to the vessel's breadth and length. The distance between vessels is determined by calculating the distance between each corner-edge pair in two bounding boxes and choosing the minimum of these 32 values. The turfpy point-to-line function was used to calculate a geodesic distance from point to line [39]. It is efficient and accurate. Given three latitude and longitude points function will return a very accurate distance from the point to the line. The function was examined on a variety of small scale objects with known distances and produced satisfactory results.

### 3.2.3 Interaction analysis

In order to determine how many tugs assisted the ship during the event, a set of rules have been made to find out whether tug was interacting with the ship.

The first rule excluded vessels docked at piers from the analysis. SOG was used to determine if a vessel was docked at a pier or not. Observations in which SOG was below a given threshold were labeled as "docked". A vessel transmits AIS messages continuously, including during the times when it is docked at a pier. Occasionally, though, the SOG of a vessel was greater than zero even when docked at a pier. That is why the SOG threshold was set to 0.11. This threshold successfully ruled out most of the cases where a vessel was wrongfully regarded as moving while leaving a sufficient margin for slower operations where the SOG of a vessel was low for longer periods of time.

The second rule assured that during the event, the distance between the tug and the ship must have been at most 100 meters, excluding vessels from the analysis that were far away. This rule was set to improve the computation time of the data processing pipeline.

The third rule avoided cases where an event was labeled as a tug operation, but in fact, the tug and the ship were just passing by at a very close distance to each other. This was done by setting the threshold for the minimum duration of the event to four minutes. However, during making the analysis, initial results showed that there were cases where the duration of tug operation could be below the given threshold. This threshold could not be shortened, because this would not rule out cases where ships were just passing by. That is why a new function determining whether the vessels were interacting and not just passing by each other needed to be implemented. This function was implemented by adding a view area to the front of the tug. Whenever the ship polygon crossed the view area of the tug at least three times during the duration of the event, the operation was labeled as a possible towing operation. Figure 3.3 gives intuition on how this function worked. The interpolated path of the vessel is just an estimation of the true path vessel has taken. From Chapter 3.2.1 it is known that the accuracy depends on the speed and the time spacing between observations. Even if the third rule was broken a handful of times, the lack of rigorous restrictions provided a strong indicator of whether the tug was assisting the ship or not.

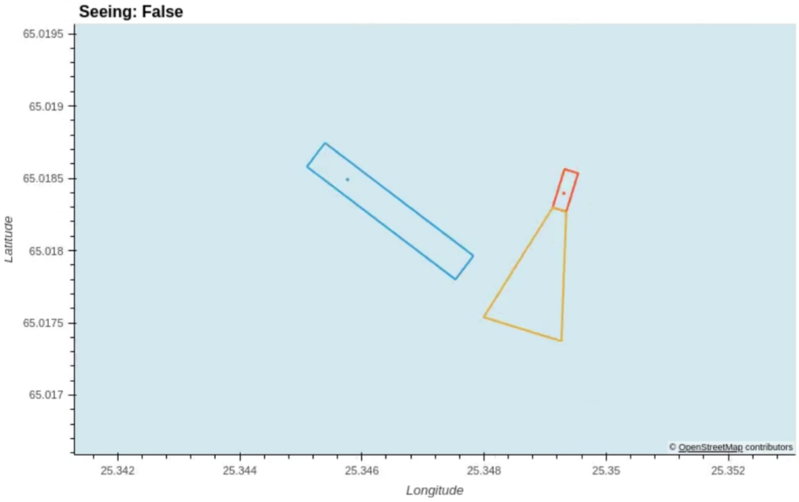


Figure 3.3: Visualization of the function determining whether the vessels are interacting. The blue ship is moving nearby the red tug. The location of the AIS transceiver is marked with a small dot in the middle of the vessel. The view area of the tug is visualized as a yellow polygon. In this case, the ship does not intersect the view area of the tug, thus the observation is not interpreted as a possible towing event. Whenever the ship intersects the tug view area three times during the event, the event will be labeled as a possible towing operation.

### 3.2.4 Finalizing the dataset

Up to this point, data is a set of events, with each event containing $n$ amount of observations. Each event represents a vessel coming or departing from the port that required or did not require tug assistance. To produce the final dataset for the learners, each event was narrowed down to a single observation. The voyage's first and last locations are the event's first and last known positions.

The type of voyage was assessed using the boundary of the port, which is a circle with a three-kilometer radius, and the center of the circle is the middle-point of the port. These boundaries cover well the ports of Rauma and Vuosaari, and give a lot of free space around the port. Whenever the first location of a ship is outside and the last is inside the port boundary, the ship is coming to the port. Whereas, whenever the first location of a ship is inside and the last is outside the port boundary, the ship is leaving the port. Whenever the SOG of the vessel falls below the threshold used in Section 3.2.3 rule one, or when the ship departs the analysis region, the voyage ends.

Wind speed and gusts in the final observation are the maximum values of each wind feature within one event. The direction of the wind in the final observation is the mean of the wind direction during the event. Since wind direction is a cyclical feature, calculating a mean value of wind direction arithmetically is not appropriate. The angular mean was calculated using a freely available algorithm on rosettacode.org [40]. Algorithm shown in Appendix B document.

Each wind direction mean value is then encoded to *sine* and *cosine* functions to keep the cyclical property of the feature. The same is done for the weekday and hour variables of the event.

Lastly, categorical variables such as ship type, voyage type, and code of the ports are converted into dummy variables. A dummy variable is a numeric variable that represents categorical data, such as ship type or port code. Dummy variables can only have one of two values: zero or one. One denotes the presence of the property, while zero denotes its absence in this analysis. The label of the final dataset is the maximum number of tugs involved in assisting the ship. It is an integer number from zero to two:

- label 0: no tugs assisted in the operation.

- label 1: one tug assisted in the operation.

- label 2: two or more tugs assisted in the operation.

Classes 1 and 2 are positive classes. Class 0 is a negative class.

The findings revealed that there were five instances in which three tugs assisted during the event, and one instance in which four tugs assisted during the event. However, because there is insufficient data to create a consistent pattern, the other six instances will be classified as label 2. Table 3.4 is a representation of the format of the final dataset. The total number of instances in the final dataset is 5542, out of which 4477 observations are label 0, 861 observations are label 1 and 204 observations are label 2. There are 3735 cargo ships, 1677 passenger ships, and 130 tankers. There were 2798 instances of ships departing the port and 2744 instances of ships arriving at the port.

## 3.3 Evaluating quality of the final dataset

Evaluating the quality of the data is important because inaccurate data may lead to invalid conclusions [41]. In this thesis, the final dataset was verified using the data provided by Alfons Håkans. Alfons Håkans dataset (AH) is a list of incoming and outgoing ship towing operations for a period stated in the earlier part of this chapter. Events themselves are verified, not the number of tugs involved in the towing event. The type of exact tug used in the event nor the number of the tugs cannot be assessed with the AH dataset, since there is no information about tugs involved in the event. It is important to emphasise that the AH dataset contains only events that required tug assistance, it does not contain cases where no tugs were needed. The purpose of using AH is to check if the algorithm was able to find the majority of towing events for both Rauma and Vuosaari ports.

Out of total 561 towing events detected in port of Vuosaari 525 (93.6%) were found in the AH dataset. For the remaining 36 events it does not immediately

Table 3.4: Format of the final dataset

| Variable | Example value |
|---|---|
| ship_type_60 (int) | 0 |
| ship_type_70 (int) | 0 |
| ship_type_80 (int) | 1 |
| ws_10min (float) | 7.4 |
| wg_10min (float) | 8.2 |
| length (float) | 185.0 |
| breadth (float) | 25.0 |
| draught (float) | 8.5 |
| hour_sin (float) | -0.942 |
| hour_cos (float) | -0.335 |
| weekday_sin (float) | 0.975 |
| weekday_cos (float) | -0.223 |
| wd_10min_sin (float) | 0.906 |
| wd_10min_cos (float) | -0.423 |
| coming (int) | 0 |
| leaving (int) | 1 |
| FIRAU (int) | 1 |
| FIVUO (int) | 0 |
| label - tug_amount (int) | 0 |

mean that the event is incorrectly labeled. Alfons Håkans is the main tugboat service company operating, however, it is possible that other organizations, e.g. port operator or other tug provider could have been used for some of the remaining 36 events. Knowing the MMSI number of the ship and the precise date of the event allowed visual analysis of those non-present 36 events. Visual analysis revealed that the majority of the events were indeed tug operations, with only a few probably mislabeled. Incorrectly labeled events are cases where a tug and a ship are passing by each-other very closely and unpredictably. The exact number of falsely labeled events is unspecified because some of the uncertain events might be, in reality, towing events. Moreover, it is impossible to determine the final label visually since the tug may be just briefly assisting the ship and could leave the area immediately if the event was considered completed.

In the case of port of Rauma out of total 504 detected events, 495 (98.2%) were found in the AH dataset. Nine non-present events faced the same situation as the previous, meaning that most of those events were correctly labeled, with a few exceptions where it was impossible to determine whether the event involved tugs or not.

AH data was also helpful in assessing the coverage of the final dataset. Out of total 512 port of Rauma actual towing events registered in AH data, nine (1.8%) were not present in the final dataset. Whereas, out of total 569 port of Vuosaari towing events 16 (2.8%) were not present. Initial analysis showed that in certain cases event data was completely missing or it was wrongly labeled as zero tugs, giving a proof that processing of the AIS data or the AIS data itself could be improved. The exact reason for this inaccuracy has not yet been resolved as the processing pipeline is yet to be fully tested. However, it is important to note that although the extraction of additional observations could possibly improve the final result, the lack of these additional observations will not negatively affect it in any way.

Despite the fact that there is space for improvements regarding the data quality, the number of events with an uncertain final label in the current state of the dataset is low enough and will only have a minor negative effect on the final model. Because of this fact, it has been decided that events with an uncertain final label will also be present in the final dataset.

# 4 Machine Learning Concepts

This chapter will provide the essential concepts for understanding machine learning and will describe how the machine learning prediction procedure was carried out. First, the theoretical section of this chapter will cover machine learning principles in order to help comprehend the process of prediction and its outcomes. The following section of this chapter will present the results obtained using the K-Nearest Neighbors (KNN) and XGBoost algorithms. The K-Nearest Neighbors algorithm was primarily used as an iterative tool to improve the quality of the data and data processing during the data understanding phase. The final result of this analysis will be presented using the state-of-the-art XGBoost technique. The next chapter will discuss and analyze the findings.

This section will start with the fundamentals of supervised learning and classification. Following that, there will be a discussion of training and test sets, over- and underfitting, model selection, and cross-validation procedures. Then the next section focuses on the confusion matrix and the used metrics in this thesis. Finally, this section will discuss both classification algorithms, K-Nearest Neighbors and XGBoost, in detail.

## 4.1 Supervised and Unsupervised Learning

One approach to categorize machine learning techniques is as supervised or unsupervised. When an algorithm's objective is to identify the label of an observation, this is referred to as supervised learning. Supervised learning is a machine learning task that involves the process of learning a model that maps an input to an output based on example input-output pairs [42]. Unsupervised learning is a type of machine learning in which the goal is to discover patterns in a dataset in the absence of pre-defined labels. The objective is to define the data structure. The two main methods used in unsupervised learning are principal components analysis and cluster analysis. Towing operations prediction is a supervised learning problem whose aim is to predict the number of tugs needed for port incoming and outgoing ships. The number of tugs variable is the output of the towing operations dataset, whereas all other features are input variables.

## 4.2 Classification

Classification is a subset of supervised learning in which algorithms learn from the provided data and makes classifications for new observations. The objective is to categorize new observations into one of two or more predefined categories. Most algorithms use the dataset to develop a model, which is then used to predict the unknown label, but some algorithms make predictions based on the dataset.

Classification can be divided into two distinct problems: binary and multiclass classification. Binary classification is a more straightforward problem in which observations are labeled to one of two classes, whereas multiclass classification entails labeling an observation to one of several classes [43].

Multiclass classification should not be mistaken with multi-label classification. In multi-label classification, each sample is assigned with a set of non-mutually exclusive target labels. Labels that are not mutually exclusive can occur at the same time. As an example, consider a movie's genre. A film may be about adventure, action, or fantasy all at once, or it could be about none of the above.

## 4.3 Training, Validation and Test Sets

A train set is a collection of data used to train an algorithm [44]. Based on these inputs, the model learns to correctly categorize, constructs all dependencies, predicts probable outcomes, and makes decisions. At this point, it is important to know the concepts of model selection and hyperparameters. Model selection refers to the problem of selecting "the best" model out of a set of candidate models in order to achieve the best predictive performance using the training set. A hyperparameter is a parameter whose value is used to control the model's learning process. The term "hyperparameter tuning" refers to the process of identifying the right combination of hyperparameters from the chosen hyperparameter range that allows the model to maximize it's performance. Both model selection and hyperparameters will be discussed in more detail later.

A validation set is a sample of data that is used to get the estimates of the model's performance, and then those measures are used to choose the best performing model [44]. A model's performance is measured using an estimation parameter such as accuracy.

Once a model and hyperparameters have been determined, it is time to test the model using data from the test set. It is critical that this data has not been used previously for model training nor validation, since the objective is to determine how the chosen method works on previously unexplored data. The test set cannot be used in the model training process, as it could introduce a bias to the end result.

How well a model performs depends on many factors. Among them is data quality; without good data, no good model can be built. Naturally, performance is influenced by the type of machine learning model used, such as KNN, XGBoost, or Support Vector Machines. Additionally, each model contains a distinct set of hyperparameters.

## 4.4 Overfitting and Underfitting

Machine learning models operate on certain assumptions. The simpler the model assumptions are, the greater the model bias is. The more complicated the assumptions are and the more closely the model fits the training data, the less bias there is.

Variance is a measure of how accurately our model estimates the target variable when a different test set is used. A high variance indicates that the model is less capable of adapting to new environments. Ideally, the accuracy of the model does not differ significantly between the training set and the different test sets.

A model is overfitted when the training data accuracy is significantly higher than the test set accuracy. Overfitting occurs when a model catches both the noise and the underlying pattern in the data. Overfitted models usually have high variance and low bias [45].

A model is underfitted when it is too simple for the data that has been modeled. A model is incapable of accurately capturing the underlying pattern of the data, generating a high error rate on both the training set and unseen data. Underfitting occurs when there is not sufficient data to build a reliable model or too simple model is used. Underfitted models usually have high bias and low variance [45].

An indispensable term upon discussing underfitting and overfitting is the bias-variance tradeoff. As the model complexity increases, its bias decreases while its variance increases. When the prediction error is reduced, the variance increases, and vice versa. The key is to determine the balance point at which these values are optimal for the modeling problem.

## 4.5 Model Selection

For a predictive modeling problem like in this analysis, model selection is the task of picking one final machine learning model from a group of candidate machine learning models using a dataset [46]. It is a method that may be used to models of different types, as well as to models of the same kind with varying hyperparameters. For instance, by selecting the optimal method from a pool of alternative methods such as SVM, KNN, or Random Forest, or by selecting the optimal model from a pool of identical SVM models with different hyperparameters. The used technique for model selection is cross-validation, which chooses a model via estimated generalization error. The model with least generalization error is chosen as the "best" model. The generalization error is a measure of how well machine learning model predicts result values for previously unseen data.

## 4.6 Nested Cross-Validation

The test dataset is used to estimate the error of the classifier, and it should ideally be large in size. However, when the sample size is small, the problem is that if data is divided into test-validate-train sets randomly, the results may vary quite a lot depending on how the split has been done. Cross-validation (CV) is one solution to this.

CV is a method for evaluating machine learning models and estimating how well the model will perform on an independent set of test data. CV consists of dividing a dataset into subsets and then carrying out the analysis on a training set while a test set is used to confirm the reliability of its results. The use of CV is recommended when there is little data to work with since the method uses every single instance for training and testing. CV is not compulsory when there is a lot of data to work with and one could simply use test-validation-training sets.

The CV measures the model's generalization but cannot completely eliminate bias. If hyperparameter tuning and other components of classifier training take place within the CV loop, then CV is proved to be unbiased [47]. This is referred to as a Nested-CV. A "nested" keyword implies that double cross-validation is used on each wrap. In the inner part of the Nested-CV, hyperparameter tuning is conducted by performing k-fold divisions on the folds used to train the model, while the outer part evaluates the chosen hyperparameters on the test set. When properly applied, the Nested-CV technique is a very useful tool for evaluating the performance of a model because it provides an almost unbiased estimate of the true error [47]. The Nested-CV algorithm prevents the model from overfitting and ensures that the model possesses generalization capabilities.

The process of Nested-CV is simple to understand. Figure 4.1 shows a visualization of how the Nested-CV algorithm works. To begin, the dataset is split in the outer loop. The split data training fold is then used in the inner loop, where the best model is selected using CV. In the outer loop, the performance of the selected model is tested against the test set. The whole process is repeated, depending on the number of splits in the outer loop.

Nested-CV mimics the process of selecting the model and estimates its performance. It is not used to find the optimal hyperparameters of the final model. The usefulness of Nested-CV is also dependent on the number of splits in the inner and outer loops. The objective is to use as much data as possible during the training process. However, Nested-CV may be computationally expensive in terms of processing power required, so the number of splits must be picked with an educated guess. While Nested-CV with fewer splits may produce a biased estimate of the independent test set error, Nested-CV with many splits may be computationally and time-intensive. Although the Nested-CV technique allows for the use of entire data to measure the model's performance, it is not the final model used to predict

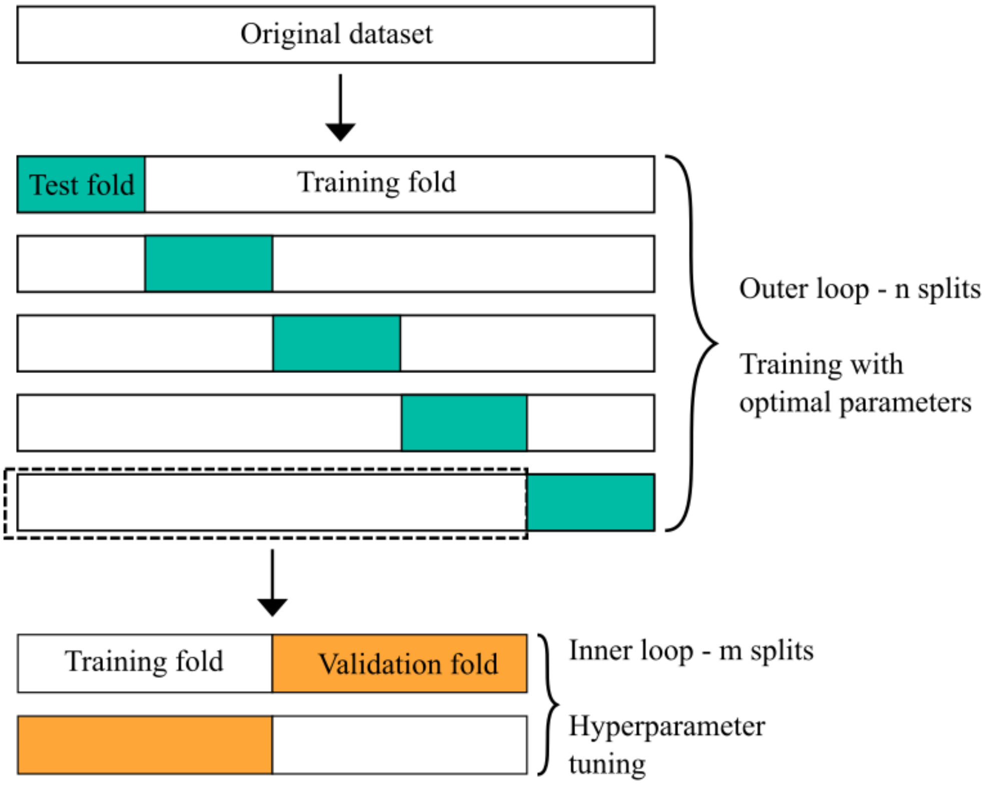


Figure 4.1: Nested Cross-Validation.

.

unseen data. The final model should be obtained by training on whole available data.

In this project, ten outer splits were used for training and testing, while three inner splits were used for hyperparameter tuning. The ten outer splits indicate that 10% of the data will be used for testing and 90% for training. The number of inner splits has a significant impact on the calculation time and has been reduced to three. These values ensure that the dataset is adequately covered, even for classes with lower frequencies. Nested-CV is completed for each model in approximately seven hours. A six-core 2.90GHz Intel Core i5-9400F processor was used, along with 16GB of RAM memory.

## 4.7 Confusion Matrix

The Confusion Matrix (CM) is the key instrument for describing the performance of the model, in other words, CM summarizes prediction result on a classification task. The CM gives insight not only into the errors produced by the classifier, but also into the sorts of errors made by the classifier [48].

CM is a two-dimensional matrix in which the rows represent the true classes and the columns represent the classifier's predicted decisions. Table 5.1 is a representation of Confusion Matrix. The number n$_{\mathrm{ij}}$ at the intersection of row i and column j represents the number of instances from the i-th class categorized into the j-th class. Chapter 3.2.4 covered the different types of classes that were utilized in this study; the total number of classes is three, which corresponds to the number of columns and rows. Model results will be presented in form of Confusion Matrices.

Table 4.1: Confusion Matrix.

| | Predicted decision classes | | | |
|---|---|---|---|---|
| Classes | Class$_1$ | Class$_2$ | ... | Class$_m$ |
| Class$_1$ | n$_{11}$ | n$_{12}$ | ... | n$_{1m}$ |
| Class$_2$ | n$_{21}$ | n$_{22}$ | ... | n$_{2m}$ |
| ... | ... | ... | ... | ... |
| Class$_m$ | n$_{m1}$ | n$_{m2}$ | ... | n$_{mm}$ |

## 4.8 Multiclass Classification Metrics

The true positive, false negative, false positive, and true negative outcomes of the confusion matrix (TP, FN, FP and TN) are utilized to calculate the metrics used in this thesis.

Precision is a metric for measuring the exactness, in other words, of the examples labeled as positive, how many are actually labeled correctly [49]. Precision can be defined as the percentage of correct predictions in positive classes (i.e. classes 1 and 2). Precision is calculated using the following formula [48]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

The recall is a metric for measuring how many examples of the positive class were labeled correctly [49]. The recall is different from precision in that it also considers missed positive predictions, not just the ones that are correct. Maximization of the precision reduces the number of false positives, whereas maximization of the recall reduces the number of false negatives [50]. Imbalanced datasets require the maximization of recall without sacrificing precision. However, this is not a trivial problem, because increasing true positives for the minority class frequently results in an increase in false positives, resulting in decreased precision. Imbalanced datasets are the datasets in which the distribution of classes is not uniform. The recall is calculated using the following formula [48]:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

To address this problem, a parameter called F-score is often used to measure overall model performance. The F-score is a metric that is used to combine accuracy and recall metric; it is used to represent both metrics with a single score. F-score is calculated using the following formula [48]:

$$\text{F-score} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \tag{4.3}$$

The last chosen metric used to measure the model performance is the balanced accuracy score, which is a multiclass classification problem metric especially defined to deal with imbalanced datasets. It is calculated as the average of recall obtained on each class. Balanced accuracy is calculated using the following formula:

$$\text{Balanced-accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{4.4}$$

Lastly, the support metric is the number of actual occurrences of the class in the dataset. Unbalanced support may suggest structural weakness in the model reported scores and may highlight the necessity for countermeasures. Support will not change between models but will diagnose the evaluation process.

## 4.9 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction algorithm, which is used to project data onto only first few principal components in order to preserve as much data's structure as possible. The first principal component can be defined as the direction that maximizes the variance of the predicted data. The second principal component is orthogonal direction to the first principal component. PCA is mostly used to decrease the number of variables used to describe a specific phenomena and to identify any patterns between the aspects. PCA makes data easier to explore and visualize, allowing more in-depth analysis of the data. A comprehensive examination of principal components allows the identification of those initial features that have a significant effect on the development of individual principal components [51].

## 4.10 K-Nearest Neighbors Algorithm

The K-Nearest Neighbors (KNN) is a simple supervised machine learning algorithm that can be used for classification and regression problems [52]. Despite its simplicity, the KNN technique produces unexpectedly good results in some applications, such as web recommendation systems [53].

KNN can be classified as a lazy algorithm, meaning that it does not contain any learning phase. This algorithm does not calculate a predictable method like linear or logistic regression. KNN measures the nearness of the data points in n-dimensional space. The nearer data points are to each other, the more similar data points are. The definition of the nearest observations comes down to minimizing a certain metric that measures the distance between the vectors of the variables of two observations. In this analysis, the Euclidean metric was used. The class of the categorized object is assigned to the class most often found among the k nearest objects.

The critical component in class selection is the parameter k, which refers to the number of nearest neighbors. K parameter may be decided adaptively or completely automatically through the use of cross-validation. In this thesis, only the k parameter will be tuned in the Nested-CV. The K-Nearest Neighbors approach is particularly advantageous when the relationship between the variables is complex or unconventional, i.e. difficult to model in a conventional manner. The relationships between the variables in our towing data are intricate and challenging to comprehend. KNN is extensively used because it is simple to implement and efficient at executing complex jobs while maintaining good classification accuracy [54]. That is why KNN was chosen above other well-known classical methods to do the initial analysis on the towing data.

## 4.11 Extreme Gradient Boosting Algorithm

Extreme Gradient Boosting (XGBoost [55]) is a popular and efficient decision tree-based machine learning algorithm suitable for classification and regression problems. A decision tree is a graph that illustrates every conceivable consequence of a decision by employing a branching strategy to represent the decision. The gradient boosting term means it is a supervised learning algorithm that predicts the target variable by combining the predictions of a set of simpler, weaker models. When adding new models, XGBoost utilizes a gradient descent approach to minimize loss.

The XGBoost algorithm was the result of a research project organized at the University of Washington. Tianqi Chen and Carlos Guestrin presented their paper at the SIGKDD conference in 2016. Since then, XGBoost has become one of the most widely used approaches in the field of machine learning due to its great results [56]. All of the trees used to develop the model have an impact on the final conclusion about the classification of individual observations. XGBoost is less difficult and time-consuming than training all trees at once. The innovation in this approach is the introduction of a regularization component. Regularization is a type of penalty placed on the model for having too many leaves in the decision tree. This is how the model's complexity is managed. As a result, the XGBoost algorithm is divided into two sections in its general version. The first component, known as the loss function or the cost function, is in charge of minimizing error. The second, regularization, prevents overtraining and regulates the model's complexity.

XGBoosting is a much more efficient version of Gradient Boosting. XGBoost uses combination of hardware and software optimization techniques to achieve good results by using minimal computing resources in a short time. XGBoost uses a gradient descent architecture to apply the boosting. Extreme Gradient Boost significantly improves the Gradient Boosting Machine backbone by implementing algorithmic improvements and system optimization. XGBoost uses a parallel implementation to start sequential tree-building process.

Figure 4.2 is illustration on how XGBoost works. In XGBoost, the trees are built sequentially. The goal of each succeeding tree is to improve on the preceding tree's mistakes. The difference between predicted and actual values in a tree is described by model residual, which is the loss incurred and will be calculated after each model predictions. The final prediction is a combination of all previous tree predictions.

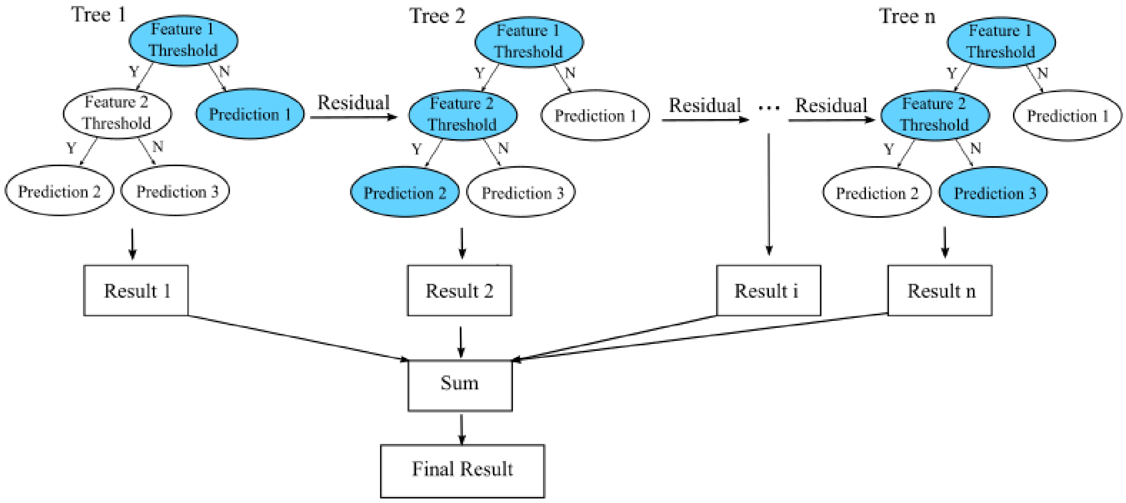


Figure 4.2: Schematic of XGBoost trees.

.

XGBoost was chosen above all other state-of-the-art techniques because decision tree-based algorithms are usually suitable for small or medium-sized tabular datasets. Additionally, decision tree-based methods like XGBoost also allows to visualize feature importance, which refers to the process of determining the relative importance of attributes in a dataset. Feature importance measures feature usefulness during the model decision tree construction. The more a decision tree uses feature to make key decisions, the greater its relative importance. The importance is calculated as the amount by which each attribute split point improves the performance measure, weighted by the number of observations handled by the node. The importance values are then averaged across all of the model's decision trees [57].

XGBoost is an extremely flexible tool that supports a diverse collection of loss functions and hyperparameters, such as the maximum depth parameter for trimming the decision tree backwards or data sub-sampling by columns and rows parameters, which prevents overfitting. The loss function used in this thesis is multi:softmax loss function.

Following XGBoost parameters have been selected for tuning:

- *max_depth*: Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. Selected values are 13, 15, 17 and 19.

- *n_estimators*: Number of gradient boosted trees. Value options were 280 and 560.

- *eta*: Boosting learning rate. Step size shrinkage used in update to prevents overfitting. Value options were 0.05 and 0.2.

- *subsample*: Subsample ratio of the training instance. Value options were 0.8, 0.9 and 1.

- *colsample_bytree*: Subsample ratio of columns when constructing each tree. Value options were 0.8, 0.9 and 1.

- *gamma*: Minimum loss reduction required to make a further partition on a leaf node of the tree. Value options were 0.0 and 5.0.

- *min_child_weight*: Minimum sum of instance weight (hessian) needed in a child. This parameter mainly helps with overfitting. Value options were 0.0 and 1.0.

# 5 Tug Operations Prediction Results

Ports were analyzed individually; one alternative was to construct a model based on a dataset including data for both Rauma and Vuosaari ports. However, learners struggled and outcomes were worse on average, despite the fact that learners had more data to work with. Each port is unique, implying that the dataset's properties are unique as well. For instance, wind blowing from north in the port of Rauma may have a different meaning than wind blowing in the same direction in the port of Vuosaari. As a result, each port was analyzed individually in order to simplify and improve the prediction process. Each port must be approached with its own set of considerations and criteria. This will be further discussed in the next Chapter 6.

The initial analysis of the Tug operations dataset was using the KNN algorithm. KNN was primarily employed as an iterative improvement process of the data processing. The data was subjected to a more advanced analysis using the XGBoost algorithm, which provided a broader set of options.

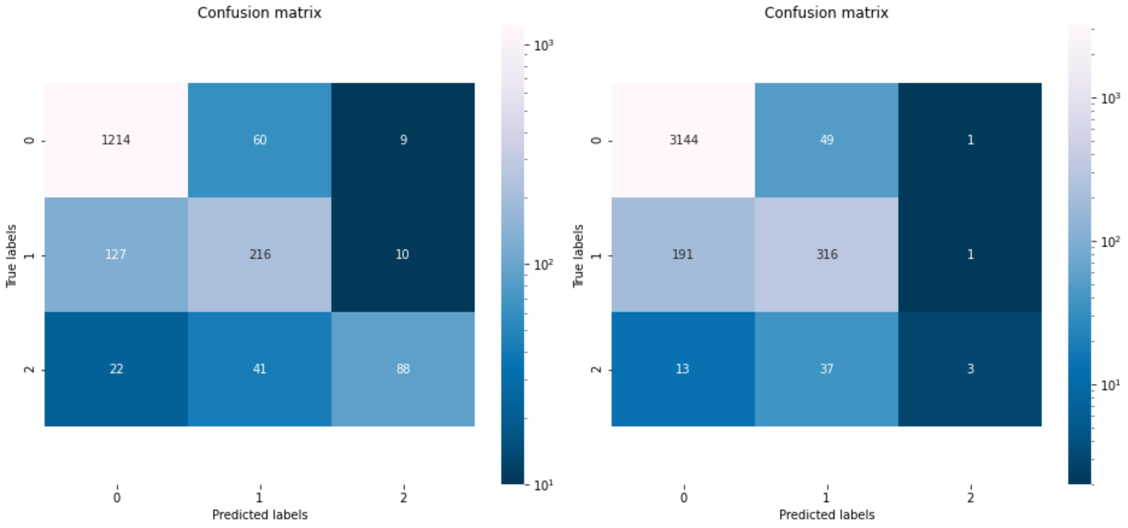
## 5.1 Initial Results using KNN

Because the used KNN algorithm distance measure is Euclidean distance, features have been standardized by removing the mean and scaling to unit variance using popular sklearn StandardScaler function [58]. This is due to the fact that KNN's distance calculation makes use of feature values. When one feature's values are greater than the others, that feature will dominate the distance and therefore the

KNN result.

The dataset classes, particularly those of the Vuosaari port, are severely imbalanced. The problem with uneven classes during detailed analysis and performance measurements of the dataset the will be handled using XGBoost weight parameters. This will be discussed in greater depth in the following section.

Figure 5.1 represents the final results obtained by the KNN algorithm for both models. KNN results are discussed in more depth in Chapter 6.1.2.



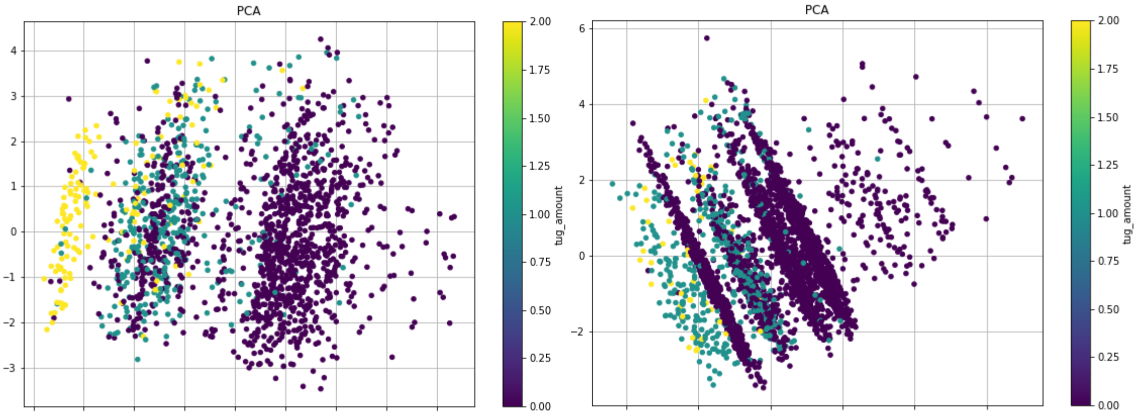
(a) Port of Rauma KNN results.

(b) Port of Vuosaari KNN results.

Figure 5.1: KNN results for both ports. It is evident that port of Rauma model predicts label 2 instances more accurately than the Port of Vuosaari model.

## 5.2 PCA

Figure 5.2 visualizes the dataset after PCA transform. Clusters are classified into three categories. The color of the cluster indicates its type. Yellow indicates data points labeled with two tugs, green represents data points labeled with one tug, and violet represents data points labeled with zero tugs. In both cases, the dataset was standardized using the same technique as described previously for KNN. PCA results are discussed in more depth in Chapter 6.1.3.



(a) Rauma dataset PCA.

(b) Vuosaari dataset PCA.

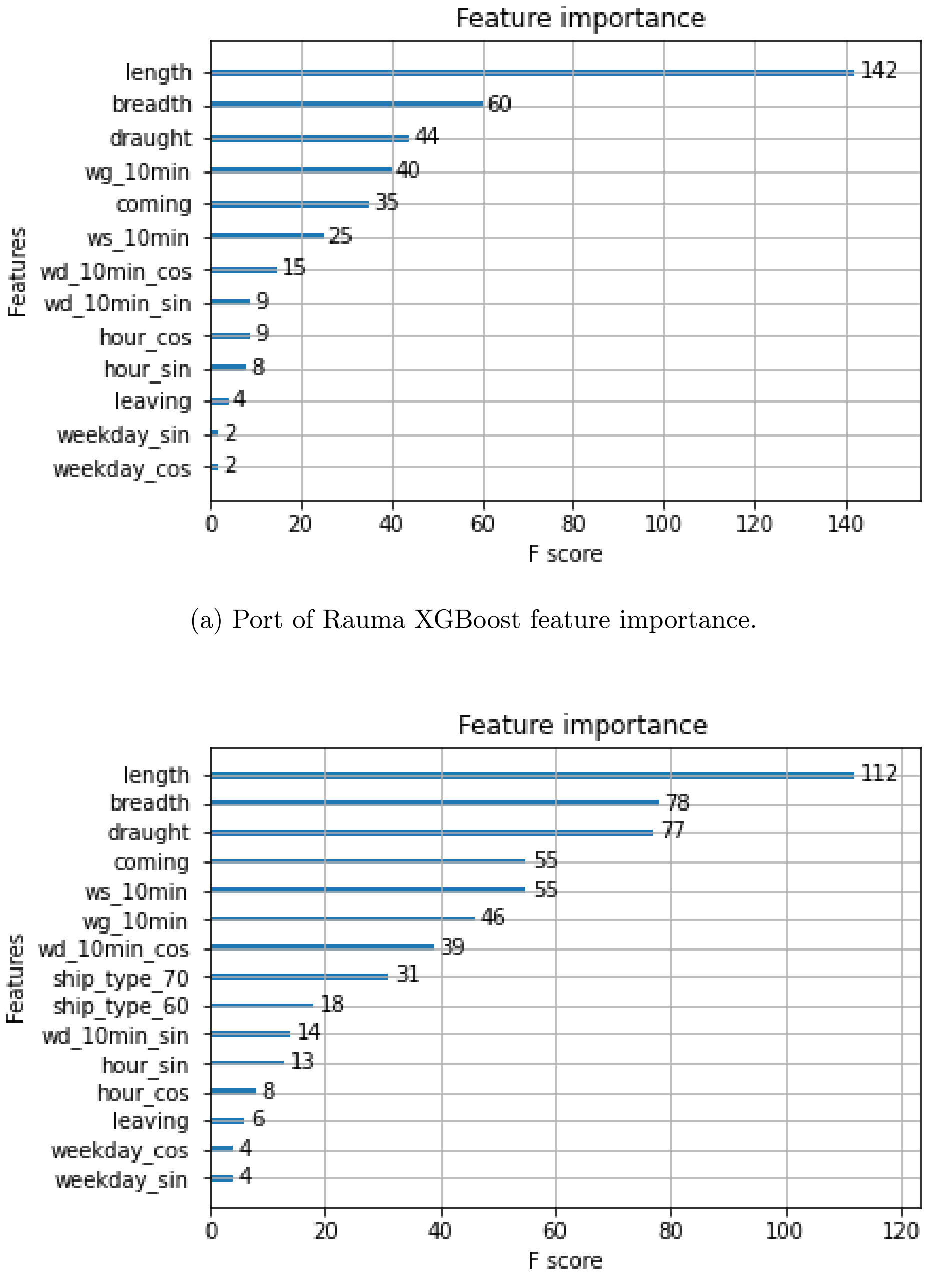
Figure 5.2: PCA of both port datasets.

## 5.3 Final Results using XGBoost

In contrast to the KNN technique, the XGBoost approach does not need feature scaling as the base learners are trees, and feature scaling has no influence on how the trees are produced.

As discussed in Chapter 4, XGBoost is capable of visualizing the relative importance of each feature. The provided scores are frequently referred to as importance scores, and in this case, the F-score is used. Each attribute is assigned a numerical value indicating its significance in modeling the problem. Figure 5.3 represents a feature importance for each port models.

When dealing with categorization problems, the issue of unbalanced data frequently arises [59]. Imbalanced classifications can lead to models with poor prediction performance, particularly for the minority class. A difficulty arises as a result of the fact that the minority class is usually more significant than the ma-



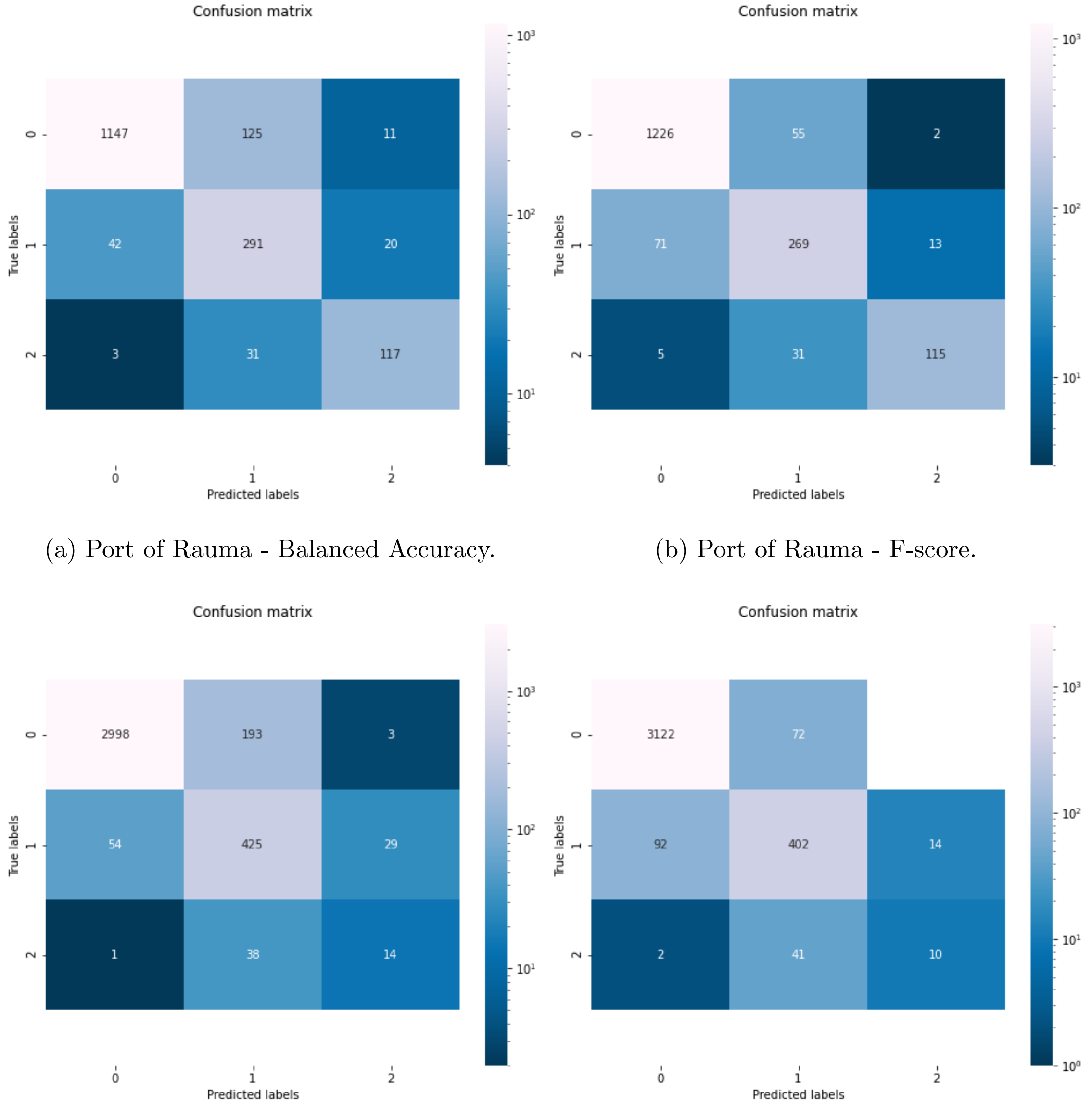
(a) Port of Rauma XGBoost feature importance.

(b) Port of Vuosaari XGBoost feature importance.

Figure 5.3: XGBoost feature importance.

jority class and, as a result, the issue is more susceptible to classification mistakes for the minority class than for the majority class. Using XGBoost, normally, the *scale_pos_weight* parameter would be used to address classes that are unbalanced, but it is only accessible for binary classification issues. In this situation, it is necessary to weight each individual data point and account for them while working with the booster, allowing the weights to be optimized automatically so that each point is represented equally. Weights are computed using sklearn compute class weight function.

Each model is computed in approximately three hours and addition of any other hyperparameters significantly increases the computation time. Selected parameters provide a good overview of the model's performance, which is sufficient at this stage of the analysis. The discussion Chapter 6 will delve into the various ways in which modeling can be enhanced. The hyperparameters have been pre-tuned separately to give the indication of the value that needs to be tuned in the Nested-CV. Figure 5.4 shows both model results in form of a confusion matrix. Tables 5.1 and 5.2 are detailed metric summaries for both models. Table 5.1 represents the results using the Balanced Accuracy metric, whereas Table 5.2 represents the results using the F-score metric. XGBoost results are discussed in more depth in Chapter 6.1.5.



(a) Port of Rauma - Balanced Accuracy.

(b) Port of Rauma - F-score.

(c) Port of Vuosaari - Balanced Accuracy.

(d) Port of Vuosaari - F-score.

Figure 5.4: XGBoost Confusion Matrix results

Table 5.1: XGBoost results using Balanced Accuracy metric.

| | Port of Rauma | | | | |
|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | F-score | Support |
| class 0 | 0.894 | 0.962 | 0.894 | 0.927 | 1283 |
| class 1 | 0.824 | 0.651 | 0.824 | 0.728 | 353 |
| class 2 | 0.775 | 0.791 | 0.775 | 0.783 | 151 |
| | Port of Vuosari | | | | |
| | Accuracy | Precision | Recall | F-score | Support |
| class 0 | 0.939 | 0.982 | 0.939 | 0.960 | 3194 |
| class 1 | 0.837 | 0.648 | 0.837 | 0.730 | 508 |
| class 2 | 0.264 | 0.304 | 0.264 | 0.283 | 53 |

Table 5.2: XGBoost results using F-score metric.

| | Port of Rauma | | | | |
|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | F-score | Support |
| class 0 | 0.956 | 0.941 | 0.956 | 0.949 | 1283 |
| class 1 | 0.762 | 0.758 | 0.762 | 0.760 | 353 |
| class 2 | 0.762 | 0.885 | 0.762 | 0.819 | 151 |
| | Port of Vuosari | | | | |
| | Accuracy | Precision | Recall | F-score | Support |
| class 0 | 0.978 | 0.971 | 0.978 | 0.974 | 3194 |
| class 1 | 0.791 | 0.781 | 0.791 | 0.786 | 508 |
| class 2 | 0.189 | 0.417 | 0.189 | 0.260 | 53 |

# 6 Discussion

In this Chapter, the major results of the thesis will be discussed. Then there will be a brief discussion of the thesis limitations, which are those aspects of the design or methodology that impacted or influenced the results. Finally, a comprehensive discussion of future work and improvements is included, outlining how to improve specific aspects of the design and outlining prospective directions for implementation projects or subsequent studies.

## 6.1 Results

This section will discuss the thesis major results and findings. To begin, there will be a brief discussion on the extraction of tugboat events from AIS data, followed by a discussion of the KNN results, which was a critical step of the project. Additionally, the KNN findings motivated to perform PCA on the dataset, which resulted in valuable discoveries about the dataset characteristics. The topic will then shift to XGBoost, beginning with a review of the feature importance analysis and concluding with a discussion of performance. The main purpose of this section is to address the following primary question: is it possible to develop a simple machine learning model for estimating the number of tugs to assist in a towing event?

### 6.1.1 Extraction of Tugboat Events from AIS Data

*The suggested method for extracting tugboat events from AIS data presented in Chapter 3 was generally successful.* The AH dataset enabled to determine that the vast majority of actual events had been discovered. However, the AH dataset did not include the number of tugboats that assisted during the event, and hence this value can be estimated and then tested. Visual inspection of extracted events revealed that while the vast majority of events were correctly labeled, several outliers were incorrectly labeled due to algorithmic flaws. Due to the large number of events, not all of them could be visually analyzed, and the real number of mislabeled observations throughout all data is unknown. As a result, the dataset almost certainly contains noise, which will affect the true performance of the model.

Although visual analysis of the extracted events has been shown to be extremely effective in identifying issues and inaccuracies during the data preparation phase, it is a time-consuming process and cannot be performed on all extracted events. Further analysis and testing of the extracted events is recommended. The following video is an example of extracted port of Rauma tugboat events with positive classes: [60].

The algorithmic inaccuracies are primarily covered in the forthcoming Section 6.2. The algorithm for extracting towing events is adequate for concluding the thesis, however, it should be improved further. Section 6.3 will examine many approaches for optimizing the data preparation process. This should theoretically result in an increase in the overall performance and accuracy of any used machine learning algorithm.

### 6.1.2 KNN Results

As previously stated, the KNN method was utilized to iteratively enhance the data preparation, data processing, and prediction due to its ease of implementation and the final results will served as a baseline for improvements during later stages of the project.

Initial analysis using KNN demonstrated that port-specific models are required. Each port is highly unique and thus requires a unique approach, meaning that one port model may not be applicable to another port. On average, the initial model for both ports produced significantly worse results than the initial model than the port-specific models. This was the primary reason for constructing port-specific XGBoost models, as it would save considerable time.

The KNN results were extremely motivating and reassuring, particularly in the case of Rauma, that is because even a simple model like KNN is capable of correctly classifying the majority of data. In principle, more complex machine learning methods such as XGBoost should be able to classify the dataset more accurately.

KNN was capable of correctly labeling the majority of the observations in classes 0 and 1, although as illustrated in Figure 5.1, the Rauma model predicts label 2 instances more accurately than the Vuosaari model, which may be due to a lack of sufficient data in the latter. However, this does not entirely explain the latter's poor performance, and additional analysis is required in that regard. PCA analysis provided very useful information for better understanding the datasets.

### 6.1.3 PCA

PCA was found to be beneficial during the data understanding phase and in determining the dataset's usability, as the clusters in the port of Rauma are more distinguishable than the clusters in the port of Vuosaari, which overlap significantly more. Therefore, machine learning algorithms should be able to predict the outcome for positive classes on the Port of Rauma dataset more easily. All factors that contribute to the quality of particular port data over others are unknown and must be investigated further.

Additionally, Figure 5.2, the far left violet long cluster is comprised of the same passenger ships that follow a constant route from Vuosaari to other ports such as Tallinn. The points are close together since the ships are virtually always identical and variation occurs as a function of the many attributes. Almost every passenger ship data point did not utilize tugs, besides the extreme circumstances of strong winds. The dataset for the Port of Rauma contained significantly less passenger ships.

### 6.1.4 Feature Importance

Figure 5.3 depicts the usefulness of features for both models. The F-score is used to quantify the usefulness of features in predicting a target output. The dimensions of the target ship were found to be the most beneficial in predicting the tug amount: *length*, *breadth* and *draught*, with *length* scoring best in both cases, as discussed in Chapter 2.3, this was anticipated. Port laws, which frequently specify the maximum allowable ship dimensions for vessels in specific port locations, are one of the most critical factor in determining the optimal number of tugs. Wind conditions features and *coming* feature, which indicated whether the target ship was approaching the port, proved also to be very beneficial. At first, it was surprising to discover that the *coming* feature received the similar score as the wind conditions features. This may imply that approaching ships require tug help more frequently than departing ships under similar conditions, and this made sense. Port operations for arriving ships are typically more complicated, as the ships may require to perform the swinging maneuver. Port operators prefer to position ships to allow for a smooth and quick departure, eliminating the necessity for swinging maneuvers, which require the employment of many tugs for larger ships. In both cases, the least useful features were *hour*, *weekday* and *leaving* features, which received the lowest scores. Features describing the hour and day of the operations were primarily introduced to simulate the availability of tugs during the weekends and during the night, although this proved ineffective and may require different approach. This might be accomplished by incorporating a traffic index feature that takes the number of moving vessels inside the port area into consideration during the towing operation.

### 6.1.5 XGBoost

This section discusses the XGBoost results and is divided into three parts. The first part discusses the XGBoost results using the Balanced Accuracy metric, the second part discusses results using the F-score metric, and the third part, comparing the two metrics used, overall discusses the results of the XGBoost and how those compare to the ones obtained by KNN.

Results for port of Rauma model shows 89.4% class 0 instances were classified correctly, 82.4% class 1 instances were classified correctly, and 77.5% class 2 instances were classified correctly. The overall accuracy is 87.0%. XGBoost results follow the same theme as the KNN results, as the model performed worse compared to the Rauma model. XGBoost did a relatively good job with instances of classes 0 and 1, but was not able to predict class 2 instances well, as previously mentioned in the KNN discussion section, this is most likely due to lower class 2 support, meaning there are only 53 class 2 observations in the Vuosaari dataset compared to 151 observations in the Rauma dataset. For the Vuosaari model, 93.9% class 0 instances were classified correctly, 83.7% class 1 instances were classified correctly, and only 26.4% class 2 instances were classified correctly. The overall accuracy is 91.5%. It is important to emphasize that regardless of the techniques mitigating the unbalanced data, the unbalance could still affect the results, meaning that the overall accuracy is biased towards the classes with higher frequency and may not give an objective picture of the overall model performance.

Accuracy using F-score for Rauma dataset negative class 0 is higher, reaching

95.6%, whereas positive classes reach a slightly lower accuracy rating of 76.2% in both cases. Here, also the port of Vuosaari results are worse compared to Rauma model results. The class 2 metrics are significantly lower, with an accuracy of only 18.9% and the model clearly struggles with capturing any pattern for this class. Class 0 accuracy was 97.8% and class 1 accuracy 79.1%.

The presented XGBoost results compare F-score and balanced accuracy metrics. Both metrics yield comparable outcomes. The F-score improves the model's overall accuracy, but at the expense of the performance of the positive classes. This is due to class 0 having a greater number of instances, implying that it is the class with the greatest influence on the final outcomes. The balanced accuracy metric maintained the relative balance of the classes and produced better results for positive classes 1 and 2. Even if a different metric system was employed, the results should not be significantly altered. There is considerably more room for improvement in other areas, such as data quality and processing. This will be covered in detail in the following Sections, 6.2 and 6.3.

Additionally, the presented XGBoost results prove that there is an improvement over the KNN baseline. The model accurately predicts the majority of instances within each class, with one exception. The majority of Vuosaari dataset class 2 instances are misclassified. As previously mentioned, this might be due to a lack of sufficient data within this class. The presented results suffice to answer the main research question and confirm that, in fact, it is possible to develop a simple machine learning model for estimating the number of tugs to assist in a towing event. Future research and implementation efforts should now be directed at improving in these areas; various directions for such enhancements are provided in Section 6.3.

## 6.2 Limitations

Perhaps the most significant limiting factor affecting the quality of the results is the absence of a variety of features capable of describing each tugboat operation event in greater detail. This would allow for a more precise description of each tugboat operation event, allowing for more accurate outcome prediction. These features may include a description of the skill set required to operate an organized fleet, as well as human resource or financial constraints. These are just a few examples of which are not available through AIS or weather data and thus require the use of additional data sources.

The AH dataset allowed to assess the detectability of events. It did not, however, allow to determine whether the number of tugs involved in the towing event was correct or not. This must also be assessed in order to fully evaluate the quality of the dataset. Visual analysis of a subset of obtained events indicated that the algorithm performed well, but it was impossible to evaluate all obtained events. Also, because the AH dataset does not include cases in which no tugs were required, the dataset contains noise, as all arriving and departing ships are considered, including those that are not within the port's operational scope. To obtain a completely noise-free dataset, the algorithm must be combined with the port-call dataset to determine whether an arriving or departing ship is within the port operational scope.

The data processing pipeline analyzed the data daily and then aggregated the results. This was done primarily to enable a parallel analysis, which allows for considerably faster calculation, rather than calculating vessel voyages one by one, which can take much longer. This solution significantly simplified the handling of AIS data, which was previously split by month due to its size. This method is also beneficial in loading data due to RAM memory constraints. However, this solution creates some complications, as certain towing events may take place at night. When a towing event begins before midnight and finishes after midnight, the event is split and treated as two separate events if the towing event's conditions are met. Finally, this could result in an incorrect estimate of the number of tugs that helped in the event. While such a mistake is possible, the chances are slim. In the port of Vuosaari, fifteen events occurred over midnight, while in the port of Rauma, eighteen such events occurred. The data processing pipeline is flexible enough to handle some outliers, and if the time window brake is small enough, the towing events are combined into one.

The current tugboat interaction analysis algorithm only considers the front view of the tugboat, implying that tugs will always face the ship at some point during the towing operation. However, visual examination of the collected towing events revealed that tugboats occasionally perform towing operations while facing the ship backwards for the duration of the event. This is especially true for contemporary tugboats, which are usually designed to operate in narrow water channels and are equipped with engines capable of operating in all directions. The preliminary analysis upon which the project's assumptions were built for the implementation phase was based on older tugboat units that always faced the ship during the towing operation.

The voyage type determination contains a conceptual error as well; for example, if an incoming ship stops within the port circle area and then proceeds, the voyage type is then incorrectly labeled as *port operation* instead of *coming*. Port operations within port area are not concern of this thesis, and hence the event would be completely overlooked meaning that the quality of the data would slightly decrease. It is uncertain how frequently this occurs, but it should be addressed in order to avoid such occurrences.

The data processing pipeline contains a minor conceptual error that may affect the interpolation of the vessel path. It is possible that whenever the vessel turns more than 180 degrees at high speeds, the direction of the turn is incorrect. That is because the AIS transceiver may not have transmitted a message between the start and end of the turn. However, it should be noted that no evidence of such error was observed during the project's visual analysis phase.

## 6.3 Improvements and Future Work

One way to enhance the model's training process is to incorporate feature selection, which is the process of choosing a subset of features to use in the model construction. There are numerous advantages to this, including the following: current modeling requires numerous hours of training, whereas a subset of core features speeds up the training and simplifies the model, making it easier to interpret. Without significant information loss, irrelevant features would be removed. Additionally, the dataset could be enhanced through feature extraction by providing new features. Throughout the process of making the thesis, I discovered that the type and power of the tugboat may play an important role in specific situations; some tugs are suited for particular sized ships, while others are not; smaller tugs lack sufficient power or are not safe to assist larger ships because of the size difference. The current model regards tugboats equally, which is not the case in reality. The power of a tugboat can be estimated using feature analysis of AIS data. Additional features defining the size and power of the tugboat may catch details not captured by the existing model, hence increasing prediction accuracy.

Because the final dataset is obtained using the processing of AIS data, it is difficult due to the size of the data to discern outliers and inaccuracies, and hence it is only an estimation of the true operations that occurred. Although the quality of the dataset used in this thesis was assessed using AH data, which was discussed in Chapter 3.3, it does not provide the precise number of tugs assisting in a towing operation. To improve the prediction of towing operations, one could obtain port-call data from the target port's officials. Port-call data is typically much more precise and includes additional features that would considerably increase the quality of the dataset, which should result in a much more accurate and exact prediction of tug operations. Collaboration with port authorities to improve data collection could increase the present set of features, such as the particular berth to which a ship is arriving or departing. Naturally, some berths are easier to reach than others and require a different number of tugs for the same ship during the same weather conditions. Most likely, a machine learning model would discover a correlation between the type of berths or terminals and the number of tugs used in the operation.

Certain ports may require a significantly more customized approach that includes an assessment of a broader set of requirements. For instance, the very busy port of Hamburg port, which is a seaport on the river Elbe. The port is located approximately 110 kilometers from the North Sea. Due to the port's unique location, it may be necessary to consider additional factors that may affect the number of tugs used in the towing operation, such as the current direction and force of the water, the level of river or port traffic, the level of visibility, the available keel clearance during the tidal window, and many more.

Section 6.2 described the possibility of incorrect interpolation of turning vessels. Although no evidence of this occurring was found, the vessel path interpolation algorithm should be improved by implementing the *rot* variable, which specifies the direction and speed of turn. *Rot* variable is available in the AIS dataset. Table A.3 contains a detailed description of the *rot* variable.

The previous section also described the issue of some types of tugs performing tugboat operations facing the ship backwards. One solution to this problem is to add a tugboat back-view area in the interaction analysis, similar to the front view area described in Chapter 3.2.3. However, the length of the back-view sides should be reduced, as vessels can bypass one another much closer to the back than to the front.

Initially, four features were defined to describe the direction of ship movement during the voyage: *coming*, *leaving*, *port_ operation* and *passing_ by*. However, only *coming* and *leaving* type voyages were required for later stages of the project. Both features can be narrowed down to a single feature simply by converting the direction features to a binary value; 0 value could mean that the ship is coming to a port and 1 value could mean that ship is leaving the port.

The decision for a three-category output was discussed in Chapter 3.2.4. In this situation, the task is a multiclass classification problem. Another possibility is to simplify the problem by classifying observations into two categories: one for which a tug is required for a towing event and another for which a tug is not required for a towing event. This would therefore be a binary classification problem. However, this approach would not reveal the number of tugs required for an event, necessitating the addition of methods for assessing the number of tugs used in the towing operation. For example, on top of binary classification, logistic regression or another more advanced approach would be required to indicate the number of tugs that assisted in the towing operation. To keep the classification process simple, a multiclass classification was employed instead. However, the combination of binary classification and logistics regression should also be taken into consideration.

The trajectory analysis of tugboat operations must be precise; in some cases, particularly in ports with complicated inland infrastructure, the vessel's interpolated trajectory may not reflect the true path. When the AIS transceiver of a quickly moving vessel sends data in a small water channel, there is a risk that the data-point of the interpolated trajectory will be in the incorrect location. It may even be inland. The existing algorithm for trajectory interpolation should be improved in the future. Numerous research has been conducted on this subject, including Sang and Wall [61], who offer a unique approach for determining the vessel route over an inland canal using AIS data. The proposed method accurately and efficiently captures a vessel's trajectory in restricted water channels.

Lastly, the presented analysis could be taken even further by taking into consideration other predictive algorithms as well. Although the XGBoost algorithm is widely used as the state-of-the-art technique for classification problems, other techniques may also be equally suitable for the objectives established earlier. One must bear in mind that there is no single "best" machine learning algorithm for all issues and so various methods, ranging from highly customized neural network-based algorithms to Support Vector Machine algorithms, should be considered as well. Further research could be conducted to find the optimal algorithm for this particular problem. It should be emphasized, however, that the gains from changing the algorithm may be small in comparison to a well-optimized XGBoost model.

# 7 Conclusion

This thesis examined the feasibility of using machine learning to estimate the number of tugs involved in a towing event. Numerous stakeholders in the maritime industry could benefit from such an automated predictive system, for example, by lowering costs through increased operating efficiency, and by lowering the risk of loss in the event of a mistake or miscommunication. Due to a lack of research on the subject, the primary objective of this thesis is to address the most critical concerns, which will help to set the stage for future research. Aiming to achieve this, a method has been proposed in order to answer two main research questions. The first part of the thesis concerns the extraction of towing events from historical AIS data, and the second concerns the development of a machine learning model used for estimating the number of tugs to assist in a towing event.

In general, the recommended ideas and their implementation were a success from a performance standpoint. The proposed method for extracting towing events was validated using historical data provided by the tugboat service provider operating in the ports analyzed; in both cases, the vast majority of towing events were successfully detected. The newly acquired dataset of towing events was then used to develop a prediction models that estimate the number of tugs required to assist in a towing event. The obtained models are port-specific. Initial analysis for the thesis showed that each port requires a unique approach by taking into consideration unique local considerations and regulations. One model solution may not work for the other.

While the predictive performance of acquired models was satisfactory in both cases, the proposed method could be further enhanced by implementing the proposed improvements and future work steps. The recommended future work steps include the incorporation of other data sources, such as port-call data from a specific port, in order to improve the quality of the acquired results. This dataset may contain extra information, such as the specific berth at which the vessel was moored. Additionally, certain ports may require a more personalized approach that takes other factors into account, such as the current volume of traffic, the direction and force of the water current, or the availability of tugboat crew.

This work demonstrates that it is possible to implement a machine learning-based predictive system for towing operations. Tugboat operations prediction system can be a significant step toward port automation, which is unavoidably the future of maritime logistics. It is only one of numerous steps that must be addressed before port operators can fully exploit the promise of AI in maritime logistics. Such a tool could enable a more efficient and safe operational management, thereby benefiting the marine logistics industry in general. Furthermore, this thesis adds to the body of literature on machine learning in context of maritime logistics because of the rising need to investigate this industry from an academic perspective due to the rapid pace of technological change. The proposed method may be utilized in future studies on this subject since the entire process can now be iterated upon and a much more advanced and efficient method can be developed.

# Acknowledgements

My thesis would not be possible without academic, professional, and personal support. I'd like to thank Dr. Petra Virjonen, my project supervisor, for her guidance and motivation during the project. Her help was vital during my research and thesis writing, and I could not have asked for a better mentor for my master's thesis. My deepest gratitude goes to Dr. Jussi Poikonen, Vaklin Angelov, and the rest of the Awake.AI team for their time and expertise on this project. Thank you for making this project a reality. Finally, I'd like to express my deepest appreciation to my family for their unwavering support.

# References

[1] P.-L. Sanchez-Gonzalez, D. Díaz-Gutiérrez, T. J. Leo, and L. R. Núñez-Rivas, "Toward digitalization of maritime transport?", *Sensors.*, vol. 19, no. 4, 2019, ISSN: 1424-8220.

[2] V. Babica, D. Sceulovs, and E. Rustenova, "Digitalization in Maritime Industry: Prospects and Pitfalls", in. Jan. 2020, pp. 20–27, ISBN: 978-3-030-39687-9. DOI: 10.1007/978-3-030-39688-6_4.

[3] J. R. Fonseca, *Valencia to Upgrade Three Ports*, Jan. 2017. [Online]. Available: https://www.marinelink.com/news/valencia-upgrade-three420380.

[4] T. M. Executive, *South Korea Plans $6 Billion in Port Redevelopment by 2030*, Dec. 2020. [Online]. Available: https://www.maritime-executive.com/article/south-korea-plans-6-billion-in-port-redevelopment-by-2030.

[5] J. Saarikoski and R. Helminen, "Satamien digitalisaation nykytila Suomessa (in Finnish)", 2019. [Online]. Available: https://www.utu.fi/sites/default/files/media/MKK/Julkaisut/B210_Satamien_digitalisaation_nykytilaselvitys.pdf (visited on 02/14/2022).

[6] S. Fan, X. Yan, J. Zhang, and J. Wang, "A review on human factors in maritime transportation using seafarers' physiological data", Aug. 2017, pp. 104–110. DOI: 10.1109/ICTIS.2017.8047751.

[7] J. Ernstsen and S. Nazir, "Human Error in Pilotage Operations", *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, vol. 12, pp. 49–56, Mar. 2018. DOI: `10.12716/1001.12.01.05`.

[8] Awake.ai, *Universal information exchange of maritime logistics.* [Online]. Available: `https://www.awake.ai/` (visited on 08/02/2021).

[9] C. Zhang, J. Bin, W. Wang, X. Peng, R. Wang, R. Halldearn, and Z. Liu, "AIS data driven general vessel destination prediction: A random forest based approach", *Transportation Research Part C: Emerging Technologies*, vol. 118, p. 102 729, 2020, ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2020.102729`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0968090X20306446`.

[10] O. Bodunov, F. Schmidt, A. Martin, A. Brito, and C. Fetzer, "Real-Time Destination and ETA Prediction for Maritime Traffic", in *Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems*, New York, NY, USA: Association for Computing Machinery, 2018, pp. 198–201, ISBN: 9781450357821. DOI: `10.1145/3210284.3220502`. [Online]. Available: `https://doi.org/10.1145/3210284.3220502`.

[11] Á. Orgaz Expósito, "Deep Learning  Graph Clustering for Maritime Logistics: Predicting Destination and Expected Time of Arrival for Vessels Across Europe", M.S. thesis, Aalto University. School of Science, 2020. [Online]. Available: `http://urn.fi/URN:NBN:fi:aalto-202008235092`.

[12] J. He, Y. Hao, and X. Wang, "An Interpretable Aid Decision-Making Model for Flag State Control Ship Detention Based on SMOTE and XGBoost", *Journal of Marine Science and Engineering*, vol. 9, no. 2, 2021, ISSN: 2077-1312. [Online]. Available: `https://www.mdpi.com/2077-1312/9/2/156`.

[13] Q. Meng, J. Weng, and S. Li, "Analysis with Automatic Identification System Data of Vessel Traffic Characteristics in the Singapore Strait", *TRANSPORTATION RESEARCH RECORD*, vol. 2426, pp. 33–43, Dec. 2014. DOI: `10.3141/2426-05`.

[14] D. Yang, L. Wu, S. Wang, H. Jia, and K. Li, "How big data enriches maritime research – a critical review of Automatic Identification System (AIS) data applications", *Transport Reviews*, vol. 39, pp. 1–19, Jul. 2019. DOI: `10.1080/01441647.2019.1649315`.

[15] V. Paulauskas and D. Paulauskas, "Research on work methods for tugs in ports", *Transport*, vol. 26, pp. 310–314, Sep. 2011. DOI: `10.3846/16484142.2011.623825`.

[16] S. Chen, F. Wang, X. Wei, Z. Tan, and H. Wang, "Analysis of Tugboat Activities using AIS Data for the Tianjin Port", *Transportation Research Record*, vol. 2674, no. 5, pp. 498–509, 2020. DOI: `10.1177/0361198120916734`. [Online]. Available: `https://doi.org/10.1177/0361198120916734`.

[17] L. Zhen, K. Wang, S. Wang, and X. Qu, "Tug scheduling for hinterland barge transport: A branch-and-price approach", *European Journal of Operational Research*, vol. 265, no. 1, pp. 119–132, 2018, ISSN: 0377-2217. DOI: `https://doi.org/10.1016/j.ejor.2017.07.063`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0377221717307063`.

[18] allaboutais.com, *AIS history*. [Online]. Available: `http://www.allaboutais.com/index.php/en/aisbasics1/ais-history` (visited on 06/30/2021).

[19] "International Convention for the Safety of Life At Sea", *International Maritime Organization (IMO)*, Nov. 1974. [Online]. Available: `https://www.refworld.org/docid/46920bf32.html` (visited on 06/18/2021).

[20] S. Wang, R. Yan, and X. Qu, "Development of a non-parametric classifier: Effective identification, algorithm, and applications in port state control for maritime transportation", *Transportation Research Part B: Methodological*, vol. 128, pp. 129–157, 2019, ISSN: 0191-2615. DOI: `https://doi.org/10.1016/j.trb.2019.07.017`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0191261519301390`.

[21] F. Natale, M. Gibin, A. Alessandrini, M. Vespe, and A. Paulrud, "Mapping Fishing Effort through AIS Data", *PLOS ONE*, vol. 10, no. 6, pp. 1–16, Jun. 2015. DOI: `10.1371/journal.pone.0130746`.

[22] M. Redoutey, E. Scotti, C. Jensen, C. Ray, and C. Claramunt, "Efficient Vessel Tracking with Accuracy Guarantees", pp. 140–151, Dec. 2008. DOI: `10.1007/978-3-540-89903-7_13`.

[23] A. Harati-Mokhtari, A. Wall, P. Brooks, and J. Wang, "Automatic Identification System (AIS): Data Reliability and Human Error Implications", *Journal of Navigation*, vol. 60, pp. 373–389, 2007.

[24] V. Paulauskas, M. Simutis, B. Plačiene, R. Barzdžiukas, M. Jonkus, and D. Paulauskas, "The Influence of Port Tugs on Improving the Navigational Safety of the Port", *Journal of Marine Science and Engineering*, vol. 9, no. 3, 2021, ISSN: 2077-1312. DOI: `10.3390/jmse9030342`. [Online]. Available: `https://www.mdpi.com/2077-1312/9/3/342`.

[25] B. Österlund, *Suomen meriliikenteen huoltovarmuudelle asetetut tavoitteet ja niiden toteutuminen (in Finnish)*. [Online]. Available: `https://urn.fi/URN:ISBN:978-951-25-3058-8` (visited on 07/01/2021).

[26] *Etusivu - Alfons Håkans*. [Online]. Available: `https://alfonshakans.fi/` (visited on 06/13/2021).

[27] UNECE, *Secretariat Note to the users of UN/LOCCODE 2020-2*. [Online]. Available: https://unece.org/sites/default/files/2020-12/2020-2%5C%20UNLOCODE%5C%20SecretariatNotes.pdf (visited on 07/01/2021).

[28] portofrauma.com, *Cargo statistics 2019*. [Online]. Available: https://portofrauma.com/wp-content/uploads/2021/01/Tavaraliikenne_eng-01.01-31.12.2020.pdf (visited on 06/30/2021).

[29] vayla.fi, *Rauma 12m fairway*. [Online]. Available: https://vayla.fi/documents/25230764/35593001/Rauman+12+m+v%C3%A4yl%C3%A4+eng.pdf/fa9033f5-7d89-4d61-bf0a-aaa8384cbccd/Rauman+12+m+v%C3%A4yl%C3%A4+eng.pdf?t=1568812907624 (visited on 06/30/2021).

[30] portofrauma.com, *Kontit*. [Online]. Available: https://portofrauma.com/palvelut/lastinkasittely/kontit/ (visited on 07/01/2021).

[31] portofhelsinki.fi, *Cargo statistics 2019*. [Online]. Available: https://vuosikertomus2019.portofhelsinki.fi/tavaraliikennelukuina/ (visited on 06/30/2021).

[32] T. Karvonen and J.-P. Jousilahti, *Helsingin Sataman vaikuttavuustutkimus 2019*, May 2019. [Online]. Available: https://www.portofhelsinki.fi/sites/default/files/attachments/Helsingin_Sataman_vaikuttavuus_2019_0.pdf (visited on 05/30/2021).

[33] vayla.fi, *Vuosaari 11m fairway*. [Online]. Available: https://vayla.fi/documents/25230764/35593001/Vaylakortti_Vuosaari_en.pdf/ce33cc4e-4a96-4495-b527-ff1f0cad3e82/Vaylakortti_Vuosaari_en.pdf?t=1445974180922 (visited on 07/01/2021).

[34] *Open data and APIs for traffic | Digitraffic - Fintraffic*. [Online]. Available: https://www.digitraffic.fi/en/ (visited on 06/13/2021).

[35] *Home - Finnish Meteorological Institute*. [Online]. Available: https://en.ilmatieteenlaitos.fi/ (visited on 06/13/2021).

[36] P. Neupane, "Detecting of Maritime Anomalies and Security, Issues Using AIS Data", 2019. [Online]. Available: https://www.utupub.fi/bitstream/handle/10024/148914/PradipNeupane_MasterThesis_validated.pdf?sequence=1&isAllowed=y (visited on 06/21/2021).

[37] T. M. S. Committee, "Adoption of New and Amended Performance Standards", 1998. [Online]. Available: https://wwwcdn.imo.org/localresources/en/OurWork/Safety/Documents/AIS/Resolution%20MSC.74(69).pdf.

[38] C. F. F. Karney, *GeographicLib*. [Online]. Available: https://geographiclib.sourceforge.io/1.52 (visited on 06/24/2021).

[39] O. Mestry, *Turfpy*, 2020. [Online]. Available: https://turfpy.readthedocs.io/en/latest/ (visited on 06/24/2021).

[40] *Averages/Mean angle*, 2021. [Online]. Available: https://rosettacode.org/wiki/Averages/Mean_angle (visited on 06/24/2021).

[41] G. A. Liebchen and M. Shepperd, "Data Sets and Data Quality in Software Engineering", *Proceedings - International Conference on Software Engineering*, PROMISE '08, pp. 39–44, 2008. DOI: 10.1145/1370788.1370799. [Online]. Available: https://doi.org/10.1145/1370788.1370799.

[42] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach.* Upper Saddle River, N.J. : Prentice Hall, 2010.

[43] S. Har-Peled, D. Roth, and D. Zimak, "Constraint Classification for Multiclass Classification and Ranking", in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, ser. NIPS'02, Cambridge, MA, USA: MIT Press, 2002, pp. 809–816.

[44] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R.* Springer, 2013.

[45] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference (2nd ed.)* New York: Springer, 2002.

[46] M. Shirangi and L. Durlofsky, "A General Method to Select Representative Models for Decision Making and Optimization under Uncertainty", *Computers Geosciences*, vol. 96, pp. 109–123, Sep. 2016. DOI: 10.1016/j.cageo.2016.08.002.

[47] S. Varma and R. Simon, "Bias in Error Estimation When Using Cross-Validation for Model Selection", *BMC bioinformatics*, vol. 7, p. 91, Feb. 2006. DOI: 10.1186/1471-2105-7-91.

[48] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness  Correlation", *Mach. Learn. Technol.*, vol. 2, pp. 37–63, Jan. 2008. [Online]. Available: https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf (visited on 02/14/2022).

[49] H. He and E. A. Garcia, "Learning from Imbalanced Data", *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. DOI: 10.1109/TKDE.2008.239.

[50] J. Brownlee, *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning.* Machine Learning Mastery, 2021.

[51] S. Mishra, U. Sarkar, S. Taraphder, S. Datta, D. Swain, R. Saikhom, S. Panda, and M. Laishram, "Principal Component Analysis", *International Journal of Livestock Research*, p. 1, Jan. 2017. DOI: 10.5455/ijlr.20170415115235.

[52] B. Mahesh, "Machine Learning Algorithms - A Review", *International Journal of Science and Research (IJSR)*, pp. 381–386, Jan. 2020, ISSN: 2319-7065.

[Online]. Available: https://www.ijsr.net/archive/v9i1/ART20203995.pdf (visited on 02/14/2022).

[53] D. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", *Applied Computing and Informatics*, vol. 12, no. 1, pp. 90–108, 2016, ISSN: 2210-8327. DOI: https://doi.org/10.1016/j.aci.2014.10.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221083271400026X.

[54] B. Šter and A. Dobnikar, "Neural networks in medical diagnosis: Comparison with other methods", *Proceedings of the International Conference on Engineering Applications of Neural Networks*, Jan. 1996.

[55] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. DOI: 10.1145/2939672.2939785. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785 (visited on 02/14/2022).

[56] T. Chen, *XGBoost - ML winning solutions*. [Online]. Available: https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions (visited on 07/16/2021).

[57] A. Giussani, *Applied Machine Learning with Python*. Bocconi University Press, 2019.

[58] scikit-learn.org, *StandardScaler*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (visited on 09/28/2021).

[59] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review", *GESTS International Transactions on Computer Science and Engineering*, vol. 30, pp. 25–36, Nov. 2005.

[60] A. Borzyszkowski, *Port of Rauma Tugboat Operations.* [Online]. Available: `https://www.youtube.com/watch?v=wZmLgKtQ9L0`.

[61] L.-z. Sang, A. Wall, Z. Mao, X.-p. Yan, and J. Wang, "A novel method for restoring the trajectory of the inland waterway ship by using AIS data", *Ocean Engineering*, vol. 110, pp. 183–194, 2015, ISSN: 0029-8018. DOI: `https://doi.org/10.1016/j.oceaneng.2015.10.021`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0029801815005582`.

# Appendix A

Table A.1: AIS update frequencies [22]

| Vessel behaviour | Time between updates | Accuracy (m) |
|---|---|---|
| Anchored | 3 min | 10 |
| Speed between 0-14 knots | 12 s | 10-95 |
| Speed at 0-14 knots and changing course | 4 s | 10-40 |
| Speed at 14-23 knots | 6 s | 55-80 |
| Speed at 14-23 knots and changing course | 2 s | 25-35 |
| Speed over 23 knots | 3 s | 45- |
| Speed over 23 knots and changing course | 2 s | 35- |

Table A.2: List of static and voyage variables contained in AIS messsages [11].

| Variable | Description |
|---|---|
| MMSI | MMSI number |
| Call sign | Craft associated with a parent vessel, should use ‘A’ followed by the last 6 digits of the MMSI of the parent vessel. Some examples of these craft include towed vessels, rescue boats,tenders, lifeboats and liferafts. |
| Name | The name of the vessel should be as shown on the station radio license. |
| Type of ship and cargo type | Number starting by digit 0 = missing, 6 = passenger, 7 = cargo, 8 = tanker, 9 = other. |
| Reference for position | Reference point for reported position. Also indicates the dimension of ship (m) |
| Type of electronic position fixing device | 0 = undefined (default), 1 = global positioning system (GPS), 2 = GNSS (GLONASS), 3 = combined GPS/-GLONASS, 4 = Loran-C, 5 = Chayka, 6 = integrated navigation system, 7 = surveyed, 8 = Galileo, 9-14 = not used, 15 = internal GNSS |
| ETA | Estimated time of arrival, may be decided by the responsible administration |
| Maximum present static draught | In 1/10m, 0 = not available = default |
| Destination | The use of this field may be decided by the responsible administration. |

Table A.3: List of dynamic variables contained in AIS messsages [11].

| Variable | Description |
|---|---|
| MMSI | MMSI number |
| Navigational status | 0 = under way using engine, 1 = at anchor, 2 = not under command, 3 = restricted manoeuvrability, 4 = constrained by her draught, 5 = moored, 6 = aground, 7 = engaged in fishing, 8 = under way sailing, 9 = reserved for future amendment of navigational status, 10 = reserved for future amendment of navigational status for ships carrying dangerous goods, 11-13 = reserved for future use, 14 = AIS-SART, 15 = not defined = default |
| ROT | 0 to +126 = turning right at up to 708° per min or higher, and 0 to –126 = turning left at up to 708° per min or higher Values between 0 and 708° per min coded by ROTAIS = 4.733 $\sqrt{ROT_{sensor}}$ degrees per min |
| SOG | Speed over ground in 1/10 knot steps (0-102.2 knots) 1 023 = not available, 1 022 = 102.2 knots or higher |
| Position Accuracy | 1 = high ($\leq$ 10 m), 0 = low (>10 m) and 0 = default |
| Longitude | Longitude in 1/10 000 min |
| Latitude | Latitude in 1/10 000 min |
| COG | Course over ground in 1/10 = (0-3599). 3600 (E10h) = not available = default. 3 601-4 095 should not be used. |
| True heading | Degrees (0-359) (511 indicates not available = default) |
| Time stamp | UTC second when the report was generated by the electronic position system (EPFS) |

# Appendix B

Rosetta.org mean angle algorithm [40]:

1. Assume all angles are on the unit circle and convert them to complex numbers expressed in real and imaginary form.

2. Compute the mean of the complex numbers.

3. Convert the complex mean to polar coordinates whereupon the phase of the complex mean is the required angular mean.