Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Utkarsh Goyal

A prediction model for cargo-specific vessel availability

Master's thesis in Applied Computer Science
Supervisor: Christopher Frantz
June 2022

NTNU
Norwegian University of
Science and Technology

MARITIME OPTIMA

Utkarsh Goyal

A prediction model for cargo-specific vessel availability

Master's thesis in Applied Computer Science
Supervisor: Christopher Frantz
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Acknowledgement

I would like to thank **Kristin Omholt-Jensen** and **Pål Robert** from **Maritime Optima AS** for allowing me to be part of Maritime Optima AS, Norway. I do not believe I would be able to investigate the maritime sector without this chance. I also want to thank them for the advice and professional evaluation they have given me throughout the thesis process.

My heartfelt gratitude goes to **Morten Omholt-Jensen** of Maritime Optima AS, who, although being quite busy with his work, made the time to listen, guide, and keep me on the right track. I am not sure where I would be without him. He kept track of my development and set up all of the necessary amenities to simplify my life. I have chosen this opportunity to thank his work gratefully.

I also want to thank my NTNU supervisor, **Christopher Frantz**, whose patience I am sure I strained to the limit. He was always so active in the whole process that the project would not have been completed without him. I'd want to thank him for all of his efforts and assistance in ensuring the success of my project. Thank you very much.

Last but not least, I would like to thank several other people who supplied useful information that aided in the effective execution of this project.

UTKARSH GOYAL

Abstract

A shortage of vessels at a port can significantly impact cargo and trade flow. Businesses can predict when and how vessels will be available by understanding the factors that influence vessel availability at a port. This knowledge can help minimize the impact of vessel shortages and ensure that cargo is delivered to its destination as quickly and efficiently as possible. Following an initial literature review of available methods, the thesis proposes a method to predict the availability of the vessels at a port specific to particular types of cargo. The technique first predicts the arrival port for all the vessels, and then it calculates the Estimated Time of Arrival (ETA) at the expected port. For the prediction, the Extreme Gradient Boosting (XGBoost) [1] model has been trained on different vessel types specifically for cargo, and for the ETA routing engine of a collaborating company of the thesis Maritime Optima AS (MO) has been used. The thesis also explores the commercial applicability of the proposed solution in different shipping industry segments by drawing on feedback from industry experts to identify opportunities as well as limitations of proposed technique. The thesis is completed in partnership with the marine firm Maritime Optima AS (MO), which offers all of the preliminary data necessary to complete the work in the thesis. Furthermore, MO has provided access to the shipping experts as required at various phases of the thesis for confirmation and validation of the results developed as part of this thesis.

Contents

Figures

Tables

Code Listings

Acronyms

AIS Automatic Identification System. xi, 1, 3, 4, 7–9, 15, 22, 25–28, 30, 33, 41, 61, 77, 84, 87, 88

CAGR Compound Annual Growth Rate. 3

COG Course Over Ground. 8, 41

DWT Deadweight Tonnage. 20

ETA Estimated Time of Arrival. v, xv, 1, 5, 9, 11, 17, 25, 28, 29, 54, 57, 58, 61, 66, 73, 81, 85, 87, 90, 94, 99

GT Gross Tonnage. 7

IMO International Maritime Organization. 7, 9, 22, 30, 33, 73, 85, 87

LNG Liquefied Natural Gas. 2, 21, 33, 37, 61, 65, 81, 84, 90

LPG Liquefied Petroleum Gas. 21, 37, 55, 58, 61, 65, 66, 81, 84

ML Machine Learning. xi, xiii, 1, 3, 5, 14, 17–20, 22–24, 26, 27, 29, 31, 34, 40, 44, 46, 49, 51, 54–56, 58, 63, 64, 66, 68, 69, 72, 74–76, 78, 85, 86, 88, 90, 92, 93

MMSI Maritime Mobile Service Identity. 8, 22, 30

MO Maritime Optima AS. v, xi, 3, 5, 11, 13–15, 17, 20, 21, 29–32, 41, 57, 61, 62, 73, 75, 77, 78, 81, 83–85, 87, 88, 91, 92, 99

MST Most Similar Trajectory. xi, 16, 17, 35, 36, 38, 87

MSTD Most Similar Trajectory's Destination. 35, 36, 38–40, 43, 62, 69, 83, 88, 89, 92

RF Random Forest. 26, 27, 51, 88

RL Reinforcement Learning. 19

ROT Rate of Turn. 8

SOG Speed Over Ground. 8

SPD Segment Path Distance. xi, 17

SSPD Symmetric Segment-Path Distance. xiii, 16, 35, 36, 38, 56, 62, 63, 68, 69, 72, 83–86, 88, 90, 92

XGBoost Extreme Gradient Boosting. v, 23, 26, 27, 51–54, 56, 63, 64, 74, 75, 81, 84, 88, 90

Glossary

AIVDM/AIVDO AIVDM contains data received from other boats in the AIS messaging protocol, and AIVDO includes data from the owner's vessel.. xiii, 9, 10, 22

UN/LOCODE The United Nations maintains a five-letter geographic classification system. The codes are issued to various locations, including ports, with the first three letters representing a nation code and the subsequent three representing a place.. 20, 43, 73, 74

Chapter 1

Introduction

1.1 Topics covered by project

The issues discussed in the thesis mostly center on the search for a technique that may identify the availability of vessels at a port for a certain kind of cargo. In addition, the focus of this thesis is also on discovering an effective method for predicting the port of arrival of the vessel. The solution to these issues will include the application of the Machine Learning (ML) algorithms on the shipping data which is Automatic Identification System (AIS) data. In addition, a collection of features that may assist in finding an improved solution for the issues described in the previous sentence will be found and evaluated. Furthermore, the thesis will contain analyses of discussions held with shipping industry professionals about the value added to the marine sector as a result of finding solutions to the issues mentioned above.

1.2 Keywords

AIS data, Estimated Time of Arrival (ETA), destination port prediction, vessel prediction at port, prediction of vessel for cargo.

1.3 Problem description

Trading in the shipping sector primarily involves charterers who own the vessels, cargo owners who want to transport the cargo, and brokers who function as a middleman between these two parties. It is commonly known that the vessel and cargo are mutually reliant; the cargo needs the vessel to convey it, and the vessel requires the cargo to stay operational. As a result, ship owners are always looking for the perfect cargo that their vessel may pick up, while cargo owners are continually looking for the best and most effective means to convey the cargo. To satisfy their requests, they contact brokers who have vessel information to advise cargo owners about vessel availability and cargo information that can be shared

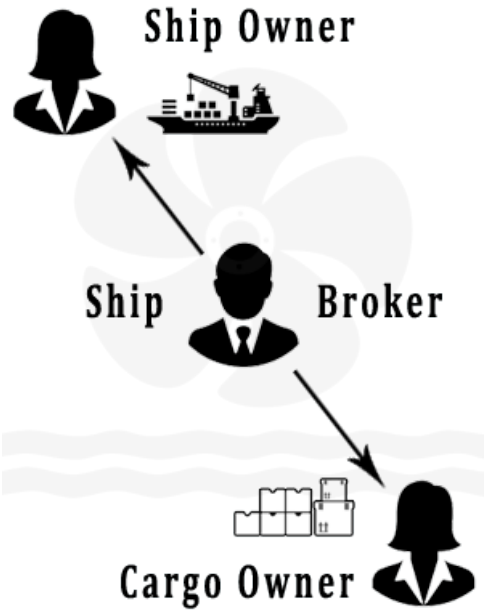


Figure 1.1: Different actors involved in the shipping industry [2]

with vessel owners so that they can prepare appropriately. Figure 1.1 highlights that the shipbroker act as a middleman between the shipowner and cargo owner.

To get information, the brokers rely on other brokers and the data from the vessel owners. For example, if the LNG cargo has to be carried, the broker will contact the various LNG vessel owners to find out where their vessels are. There is also a connection between different brokers who may offer information about other LNG vessels, but this information transmission comes at a price. The broker must offer another broker with money or additional information in exchange. As a result, all of these things are based on trust between the parties: the cargo owner will trust the broker's report, and the broker will trust the information supplied by the vessel owners and other brokers [3]. There is no procedure to validate the information provided by different parties involved.

The vessel owners want their vessels to operate efficiently and maximize their operational returns. If many vessels compete for the same cargo, the vessel with the lowest price will secure the shipping contract. As a result, vessel owners try to schedule their vessels so that there are not too many vessels fighting for a single cargo at any port. For example, if there is a chemical cargo to be picked up from Oslo next week and another from Malmö, approximately 20 boats are arriving in Oslo, and only two in Malmö. The new vessel owner will want to direct their vessel towards Malmö as there are only a few vessels arriving, so there will be less competition than Oslo. The vessel owners rely on brokers to get this information

as well. Again, vessel owners must depend on the brokers' information to ensure that their vessels sail efficiently.

Therefore there is no transparency in the existing system, and all actors are dependent on each other for the information, and there is also no method to validate the information that has been supplied between the actors. However, vessel information has been accessible in recent years through Automatic Identification System (AIS) data. However, the AIS data is mainly inconsistent and erroneous, requiring extensive pre-processing before it may be relevant. So, in recent years, several firms have worked to acquire helpful information from AIS data and assist the shipping sector. One of them is the thesis's partnering firm, Maritime Optima AS (MO), which has a large quantity of data gathered every second. As a result, this thesis makes an effort to design a system for predicting the availability of a vessel in port for a given cargo using AIS data. According to specialists, there is no publicly exist solution on the market that can supply such information. So the study done in this thesis would be beneficial for both commercial benefit and academic gain in developing many more useful solutions linked to this.

1.4 Justification, motivation and benefits

Shipping is an essential part of commerce, and it's no surprise that the shipping industry is one of the fastest-growing sectors in the economy. According to [4], the worldwide cargo shipping industry is expected to increase from 11.09 billion tons in 2021 to 13.19 billion tons in 2028 at an Compound Annual Growth Rate (CAGR) of 2.5% between 2021 and 2028. This growth is due to the expanding use of shipping services to connect buyers and sellers all over the world, and to the increasing demand for shipping goods and services. In addition to all of that, there is a massive quantity of data that is now accessible within this sector. This data is gathered daily, and it has been expanding tremendously on a daily basis. According to the website[1] for Maritime Optima AS (MO), every second, they collect AIS messages from 85 000 distinct vessels. As a result, highly important information that might be of use to the marine sector can be uncovered via the use of appropriate data analytics. On this data, numerous different ML models may be used, and a number of things can be anticipated, which can have the potential to provide enormous benefits for the industry. It is always thrilling to perform research and uncover something that may be of use to a significant number of other individuals and companies. As the shipping sector, particularly, continues to expand, with large amounts of readily available data, there remains a huge opportunity to extract valuable information that may be of strategic benefit in this industry.

This thesis has been developed with the assistance of maritime company Maritime Optima AS (MO), and reflects a continuation of the effort established in pre-

[1] https://maritimeoptima.com/

vious research performed with the same company [5]. After reading the previous research, there was a drive that with more research in the field of the maritime industry, the current thesis work can improve and expand selected aspects identified in earlier research in order to create further commercial benefit.

1.5 Research questions

This thesis made an effort to enhance and broaden the previous research solution that had been completed in the past, as mentioned in the Section 1.4 of this thesis. As a result of this, there were two goals that were established for the thesis, namely:

1. To study the past models used to predict the future destination using the AIS data.
2. To predict availability of vessel at a port in future for a particular type of cargo.

For the first objective, in this thesis there will be an analysis on what are the different models and the feature set which are been used in the past to solve the problem. After the study of different models and feature set which have been used, in the end there will be an attempt to reach to a better accuracy of the previously developed model, either by using more features or improving the model.

The second aim of the thesis is basically a reverse scenario of the previously aforementioned situation, as in this instance a port is supplied now the prediction have to be done on the vessels which can arrive at a port. For this scenario research will be done to investigate if there is any prior study accessible try to followed by an attempt to address this issue, and subsequent discussion of the results with shipping experts.

The objectives have been subdivided to form the Research Questions which are described as follows:

1. How can the AIS data be used to predict the future destination of the vessel?
    a. What kinds of models and data have been used to predict the destination of the vessel?
    b. What additional features can be added to improve the performance of the existing models?
2. What are the methods by which prediction of the availability of vessels at a port that carry specific cargo can be made?
    a. What kind of research methods have been used to predict the availability of vessels at a port for a specific cargo?

b. If previous approaches are limited, what approach could be used to predict the vessels at a port for a specific cargo?
c. To what extent can this prediction be of practical value for the maritime industry?

1.6 Planned contribution

The primary contribution of this thesis is a method for determining the availability of vessels at a port for a particular cargo. This strategy consists of two components: first, the prediction of the port of arrival, and second, the calculation of the Estimated Time of Arrival (ETA) to the projected port of arrival. Both of these techniques pose enormous challenges in their way. The collaborating organization of a thesis Maritime Optima AS (MO) has developed a tool for calculating ETA to any port in the most efficient manner feasible.Therefore, main research is conducted to identify the current solution for the prediction of arrival; based on this study, enhancements are made to the existing solutions in order to get better results. The thesis also includes the unique way of finding the arrival ports for the vessels, for certain vessels, the arrival port has been projected using the trajectory similarity approach, while for others, the arrival port has been predicted using the ML model. Finally, the concept is examined with shipping industry professionals for validation and to identify maritime sector use cases.

1.7 Remaining thesis structure

The rest of thesis consists of following parts:

- **2- Background**: In background section all the concepts, terminologies are been explained which are essential for the clear understanding of the thesis. It also include overview of the database and the challenges which arrive during the development of thesis.
- **3- Related Work**: Related work section include all the research findings from the past study to find out the research which already have been done in the related areas.
- **4- Methodology**: In methodology section all the steps which are been discussed in detail that are been followed during the creation of solution in the thesis.
- **5- Results**: Result section include the results of every processes that have been followed during the development phase. It also includes the comments from the experts and the validation of the proposed solution.
- **6- Discussion**: Discussion section finally summarize the thesis, followed by the application of the solution and the possible answers to the research questions. In the end it is concluded by stating the limitations and future work with concluding statement.

Chapter 2

Background

This chapter will go through the topics that will aid in the comprehension of the thesis work. It will cover the terminology, tools, and programming languages used for development, a quick overview of the database and its specific characteristics, alongside some challenges that will be addressed in later portions of the thesis.

2.1 Terminologies

In this section, all the essential and relevant terminology to the thesis will be discussed to understand future concepts easily.

2.1.1 Automatic Identification Systems (AIS) data

The Automatic Identification System (AIS) is used as an automated transmission system that transmits navigational signals from the vessels which have AIS transponder to the AIS receiving stations. It is extensively used in the marine industry to solve many problems such as collision avoidance, fishing, and security issues. Marine organizations also use it to perform data analytics and gain economic gains. Figure 2.1 shows the transmission of data from the vessels that have AIS transponder which transfers the data to the AIS receivers through the satellite. Finally, it is transferred to the data users.

Since December 2004, the International Maritime Organization (IMO) has required all passenger and commercial vessels traveling internationally and weighing over 299 Gross Tonnage (GT) to carry a Class 'A' AIS transponder (which transmits and receives AIS data), while smaller vessels will be equipped with a Class 'B' AIS transponder.

One sort of information communicated by AIS is dynamic, while the other is static. The data is transferred irrespective of whether the vessel is moving or anchored. The features of both dynamic and static information in the context of

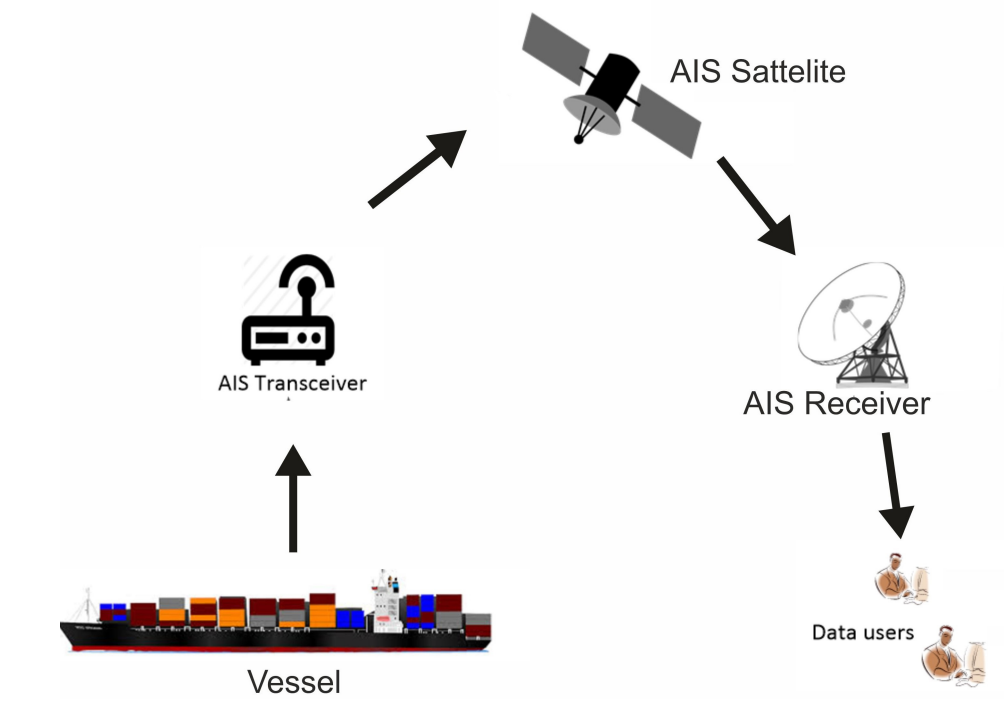


Figure 2.1: Transmission of AIS data from the vessel to the data users

AIS is explained below.

- **Dynamic Data**:
  Dynamic information is transferred every 2 to 10 seconds; it usually depends on the vessel's speed and course while moving. If the vessel is anchored, the information is transferred every 6 minutes. The following fields are included in the dynamic data:

  - **Maritime Mobile Service Identity (MMSI)**: A unique identification number for the vessel station.
  - **AIS Navigational Status**: These are the codes that tell the status of the vessel, and the crew manually sets it. The navigation code has been transmitted with every AIS message until a crew member changes it. Table 2.1 shows all the possible navigational statuses.
  - **Rate of Turn (ROT)**: It is the information of the turning (rotational) speed with the direction, right or left (0 to 720 degrees per minute).
  - **Speed Over Ground (SOG)**: It is the information of the current speed of the vessel in Knots.
  - **Position Coordinates**: This specifies the exact location of the vessel by specifying the latitude and longitude of the vessel.
  - **Course Over Ground (COG)**: It specifies the direction of the vessel

heading relative to the land.
  - **Heading**: It is the direction vessel is pointing at the given moment (0 to 359 degrees).
  - **Bearing at own position**: A bearing is a relative direction measure form the accepted reference line. It is measured in degrees.
  - **UTC Seconds**: This is the UNIX time stamp of information transmission.

- **Static Data**:
  This is the information that is manually entered by the vessel crew, and it is transmitted every 6 minutes irrespective of the vessel movement status.
  - **International Maritime Organization (IMO)**: It is the unique number of the vessel and it never changes, even in the case of vessel registration to another country.
  - **Call Sign**: The vessel's country of registry has assigned it an international radio call sign.
  - **Name**: The vessel's name, which can be up to 20 characters long.
  - **Type**: It specifies the type of vessel, or the type of cargo it is carrying.
  - **Dimensions**: It gives the dimensions of the vessel to nearest meter.
  - **Location of the positioning system's antenna on board the vessel**: It is the information of the turning speed with the direction right or left (0 to 720 degrees per minute).
  - **Type of Positioning System**: It specify which type of positioning system vessel is using. Some examples of such types are: GPS, DGPS, Loran-C.
  - **Draught**: It is the vertical distance between waterline and lowest part of vessel. It is measured in meters.
  - **Destination**: It is the name of the port where the ship is expected to arrive. However, it is frequently empty or wrongly entered by the crew. In studies it is reported not to be entered 62% of times by crew members [6] and only 4% of times its values are correct [7].
  - **Estimated Time of Arrival (ETA)**: It is the moment when the vessel will arrive at the destination port. However, in the majority of situations, this information is either incomplete or inaccurate.

All of these AIS messages follow the AIVDM/AIVDO[1] protocol, which means all the AIS messages are encoded when transmitted from the vessel due to security reasons and can only be decoded using the instructions given in the AIVDM/AIVDO protocol.

2.1.2 Haversine distance

Haversine distance is used to calculate the distance between any two points on a sphere. It is the most frequently used formula to calculate the distance between

[1] https://gpsd.gitlab.io/gpsd/AIVDM.html

Status	Description
0	Under way using engine
1	At anchor
2	Not under command
3	Restricted manoeuverability
4	Constrained by her draught
5	Moored
6	Aground
7	Engaged in Fishing
8	Under way sailing
9–13	Reserved for future use
14	AIS-SART is active
15	Not defined (default)

Table 2.1: Navigational statuses in the AIVDM/AIVDO protocol.

two geographical points on earth. In addition, it calculates the great circle distance, the shortest distance between any two given points on the sphere. The Haversine distance will be used to calculate the distance between the vessel and the port in the thesis.

Equation (2.1) shows the Haversine distance calculation formula between the two geographical points on earth, 'A' and 'B' as shown in Figure 2.2.

$$d = 2R\ \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\psi_2-\psi_1}{2}\right)+\cos(\psi_1)\cos(\psi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \quad (2.1)$$

where:

$\psi_1\ and\ \psi_2$ = are the latitude of point A and point B
$\lambda_1\ and\ \lambda_2$ = are the longitude of point A and point B
R = Radius of earth which is 6371Km

2.1.3 Douglas Peucker algorithm

The Douglas Peucker algorithm is used to make simplified polygons with fewer points than the original, also maintaining the actual shape of the polygon. The procedure begins with a basic simplification of the original polyline, a single edge connecting the initial and end vertices. The distance between all intermediate vertices and the edge is then calculated. The vertex furthest from that edge, with an estimated distance more significant than a given tolerance, will be tagged as a

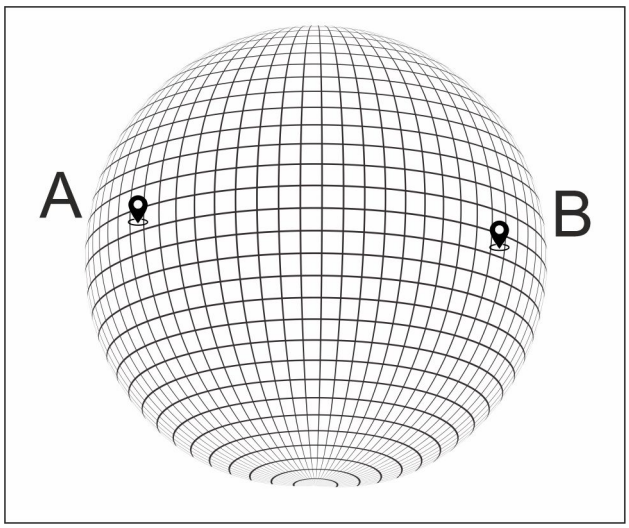


Figure 2.2: 'A' and 'B' two geographical point on earth to calculate haversine distance

key and included in the simplification. This method will repeat each edge in the current simplification until all vertices of the original polyline fall inside the simplification findings' tolerance.

Figure 2.3 shows the process; at first, the simplification is limited to a single edge. The fourth vertex is designated as a key in the first phase, and the simplification is changed accordingly. The current simplification's initial edge is handled in the second phase. No new key is inserted because the maximum vertex distance to that edge is less than the tolerance level. In the third stage, a key for the present simplification's second edge is discovered. The simplification is updated after this edge is divided at the key. This method is repeated until there are no more keys to be found. It's worth noting that just one edge of the current simplification is processed at each stage. Finally, the line in green color is the sampled polygon after the full process of the Douglas Peucker algorithm [8].

2.1.4 Tools and languages

For the construction of the model in the thesis, two main programming languages were used. One of them is **Go**[2], which is primarily used to create tracks for the voyages, and it is also used to calculate Estimated Time of Arrival (ETA) for predicted ports. **Python**[3] is the second language, which is primarily used for constructing machine learning models and performing some calculations on the dataset before giving it to the machine learning model. Go has been used because the routing engine provided by MO has been used to calculate ETA has been written in Go, so it

[2]https://go.dev/
[3]https://www.python.org/

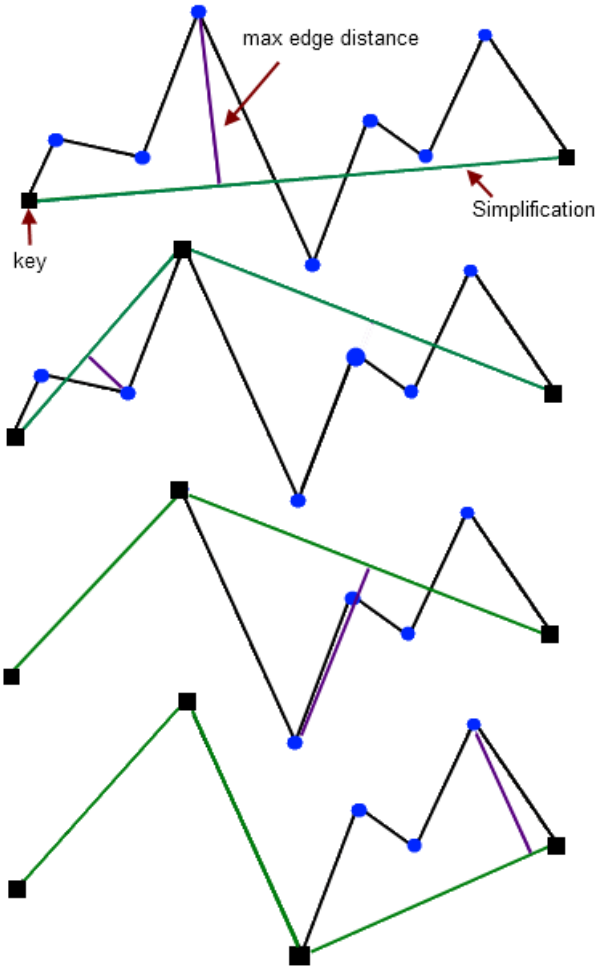


Figure 2.3: Sampled polygon after Douglas Peucker algorithm [8]

was preferred to write the new code in Go only to interact with the MO's routing engine. Python has been used because it has a wide variety of frameworks and packages for building machine learning models, making it the ideal choice for machine learning. Python also contains predefined libraries for performing calculations on geo-spatial data like *traj-dist*[4] which contain methods for comparing two trajectories, making it a better choice for performing calculations on a dataset.

The complete data which has been used in the thesis is stored in the **PostgreSQL**[5] database. PostgreSQL, also known as Postgres, is a powerful, accessible, and open-source relational database with some extended features respective to SQL database. One of the main reasons to use Postgres as a database is because of its ability to handle geo-spatial data and geometric data. Postgres achieves this with the help of **PostGIS** extension. It is easy to work with geographical data without having to convert it from the format used by the rest of the application to the one used by a database with the help of the PostGIS extension. PostGIS can also be used for data visualizations. So PostgreSQL is seemed to be an ideal choice for the database. Therefore, when referring to this thesis's suggested approach and outcomes, terminologies like database, table, row, and column refer to the PostgreSQL database and its tables with rows and columns.

Some of the other tools which are being in the thesis are described below:

- **Qgis**[6]: This tool has been used for the visualization of geo-spatial such as trajectories for the vessels, all the images in the thesis described with the map are taken with the help of this tool.
- **Google Colab**[7]: For running the machine learning model, Google Colab and its resources have been used.
- **Azure Cloud**[8]: Virtual Machine has been provided by Maritime Optima AS (MO) for performing operations with the database; due to the large size of the database, high computing power was required. So the cloud provided the virtual machine with a Linux terminal and has 256GB of Ram and 1TB of storage to save the data.
- **pgAdmin 4**[9]: This is the IDE for performing the PostgreSQL queries, it's very convenient to use, and geo-spatial data can also be visualized using this IDE.

Following the introduction of technology used for this process, the following section will turn to concepts central to this work.

[4]https://pypi.org/project/traj-dist/
[5]https://www.postgresql.org/
[6]https://qgis.org/en/site/
[7]https://colab.research.google.com/
[8]https://azure.microsoft.com/en-us/
[9]https://www.pgadmin.org/

2.2 Concepts

This section will define the concepts that have been used for the thesis solution. This part aims to provide the reader with a basic knowledge of the underlying principles that the thesis will later allude to.

2.2.1 Voyage definition

Voyage is defined for a vessel in a manner that from which port the vessel departed and to which port the vessel arrived. For the voyage definition, the departure port and arrival port should be the loading and unloading ports for the vessel. This is difficult to decide, especially for the large vessels. Larger vessels, for example, commonly bunker (refuel) in bunker ports between their journeys, vessels may dock outside of bunker ports awaiting refueling by bunker vessels, or they may slow their speed and be refueled without ever stopping altogether. Congestion in ports is another usual cause for vessels to halt moving physically. Vessels of all sizes usually have to wait for them to load or unload at crowded ports. Vessels commonly have to wait for passage through narrow canals, such as when traveling through the Suez canal, which has a small passage so only one large vessel can pass at one time. While they wait for access, they may anchor closer to a different port than the arrival port.

Predicting the ports of vessels other than the vessel's actual arrival port, where the vessel is loaded or discharged, is regarded as worthless in the maritime sector. As a result, while establishing the criteria for the vessel's voyage, only the actual departure and arrival ports should be considered; any other ports where the vessel may have stopped should be excluded. Furthermore, it is critical to have the right database for the voyage since the ML model learns on existing data patterns. Therefore, if the model learned from inaccurate data, it would provide incorrect results. Furthermore, the ML model in this thesis is heavily dependent on the voyage, and the path vessel took throughout the voyage. Hence it is critical to have the actual ports for the vessels' voyage data..

In the thesis, the database for the voyage has been used, which has been given by the Maritime Optima AS (MO). MO have their definition for defining the vessels' voyage. For checking the vessel's arrival and departure at a port MO have designed the polygons around the port so if the vessel is in the polygon, the vessel is considered as 'arrived.' If the vessel is leaving the polygon, the vessel is considered as 'departed.' Maritime Optima AS (MO) have defined the polygons around each port with a radius of 25 km keeping port location as the center and also created polygons for the berths that exist in the port. So there are two sets of polygons, one for the berth and another around the port. For both of them, there is different definition to mark the vessels' arrival and departure. If the vessel has arrived at any of the berth polygons and its speed becomes zero, then the vessel

is considered to be 'arrived,' and if the vessel moves out from the berth polygon, it is considered to be 'departed.' For the polygon around the port, if the vessel reached inside the polygon and the navigational status as defined in Section 2.1.1 becomes 'MOORED,' then the vessel is considered to have arrived at the port, and for the departure if the vessel moves out of the polygon also the navigational status changed to 'UNDERWAY SAILING' then the vessel is considered to be 'departed.' This definition has been tested by the different shipping experts and has also been used by the Maritime Optima AS (MO) in their software, so for this thesis also, this definition has been used to define the voyage for the vessels.

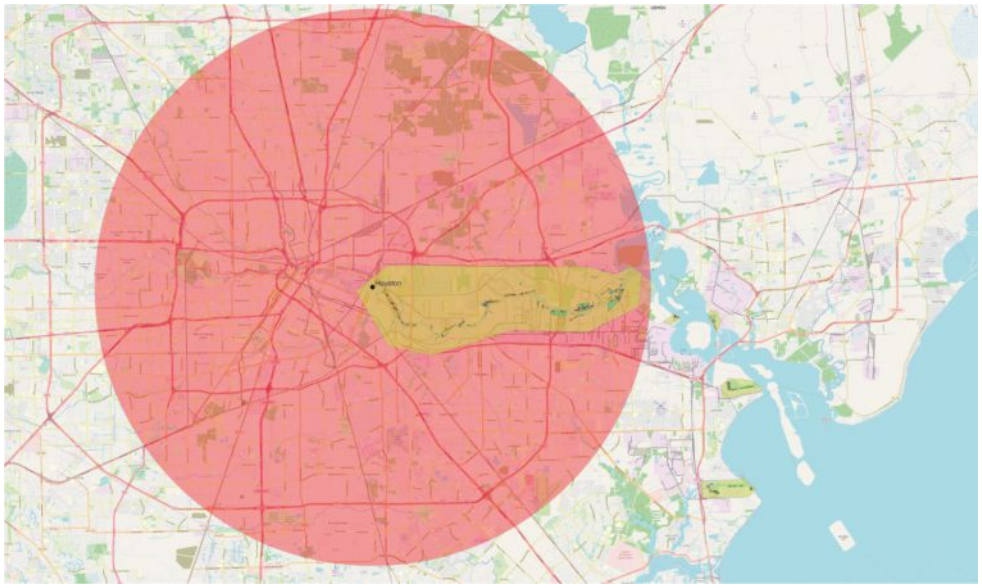


Figure 2.4: 25Km Radius around a port

The Figure 2.4 shows the 25Km radius around the port of Houston, and the Figure 2.5 is the magnified image for the berth polygons that have been designed manually by the experts of MO. So, according to the definition of the voyage, if the vessel arrived at any of the berth polygons and speed becomes zero, it is considered 'arrived.' But due to the loss of AIS signals from the vessel, the vessel position cannot be determined in any of the berth polygons. But, the vessel had reached the 25Km radius polygon before signal lost, and the navigation status was also 'MOORED' than also vessel is considered 'arrived.'

2.2.2 Trajectory similarity

Vessels' are likely to follow known shipping routes or the most economical and fuel-efficient path instead of taking new or uncommon routes for the same voyage, so the present trajectory appears to give a good insight into their ultimate destination by comparing to historical trajectories. As a result, the most similar trajectory destination port to the present moving trajectory can be determined

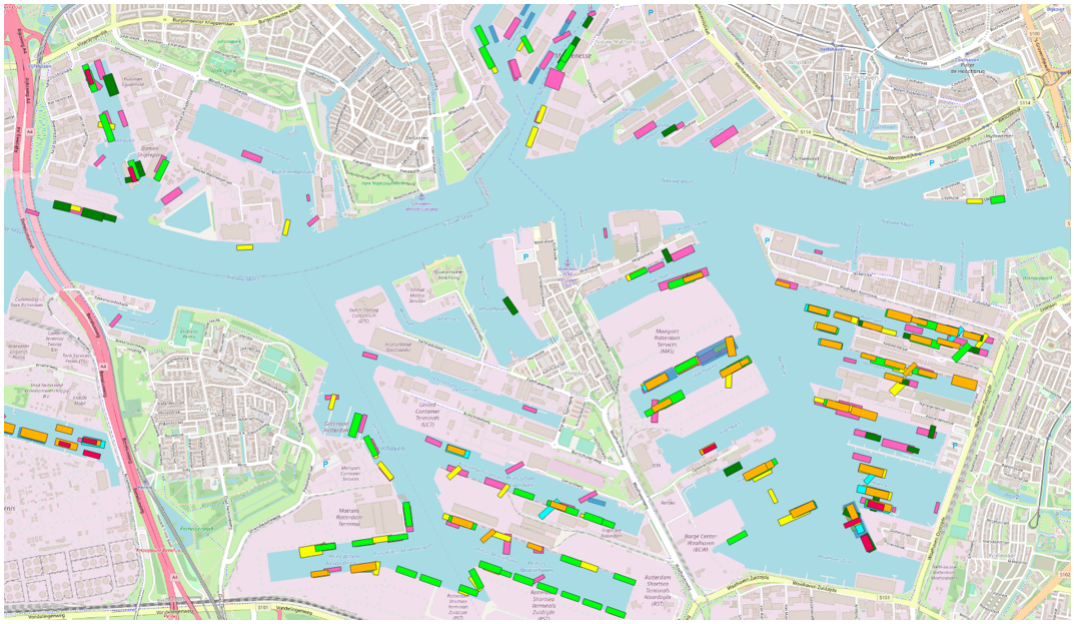


Figure 2.5: Magnified image of the berth polygons

from historical trajectory data. Therefore, the destination port of the most similar trajectory can be the first guess for the arrival port of the vessel. Through this feature, there will already be a port prediction on top of that machine learning model can be applied to improve the results.

The trajectory similarity can be of three types: spatial, temporal, and tempo-spatial. But in the marine industry, the vessels always depart and arrive at different times for the same journeys because the vessels can have different speeds or have more waiting time at canals or bunker ports. Therefore, only a spatial trajectory has been considered for the vessel's trajectory.

Symmetric Segment-Path Distance (SSPD) is a purely spatial similarity measurement method for comparing geometric shapes between trajectories that are not constrained by the length of the trajectories. SSPD compares trajectories as a whole. Therefore it's less influenced by little differences between them. In addition, the overall length, variation, and physical distance between two trajectories are all taken into account by SSPD. Figure 2.6 shows the process where the distance between points of trajectories has been calculated to find the Most Similar Trajectory (MST). The trajectory with the shortest distance will be selected as the MST.

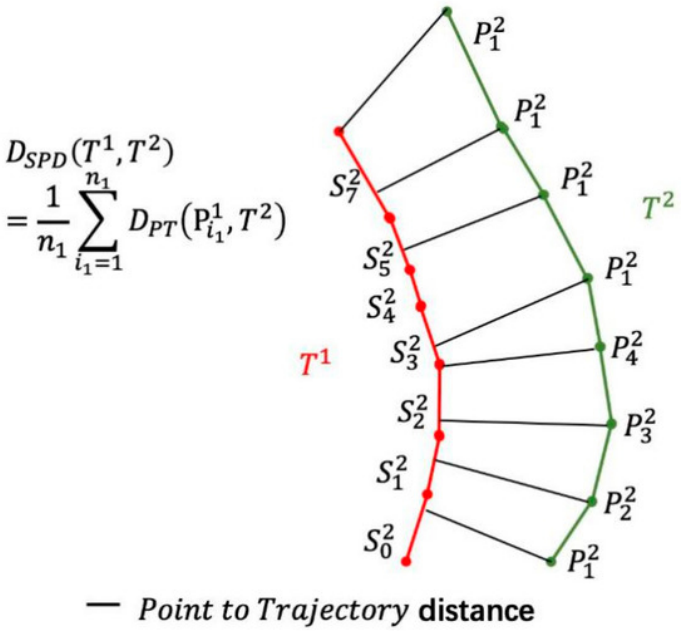


Figure 2.6: Segment Path Distance (SPD) is calculated between two points to find Most Similar Trajectory (MST) [9]

2.2.3 Routing engine

Maritime Optima AS (MO)'s routing engine is a triangulation-based path finding algorithm, a well-known path finding technique used in robotics and in computer games. Most of the core ideas are well described in the master thesis [10]. It avoids traversing close to land by shrinking the entry point between triangles, and by penalising routes that walk too close to land. It is also aware of canals such as the Suez and Panama canal, and forbids vessels to route through the canal if they are too big. In this thesis, routing engine is been used for calculating the Estimated Time of Arrival (ETA) to the predicted port.

Figure 2.7 shows the result of the route planner for the vessel 'JONAS OLDENDORFF,' the time is taken by it to reach 'Bhavnagar,' India, from its current position. The estimated route taken by the vessel is shown and the time taken to reach the destination with other information such as distance covered and the total consumption that will be made during the voyage.

2.2.4 Machine Learning(ML)

The topic Machine Learning (ML) focuses on making the computer systems learn from data by detecting "patterns" within the data. ML is a technique that has been around for a while but is currently gaining popularity because of the low cost of hardware, computation, cloud technologies, storage, and the growing number of

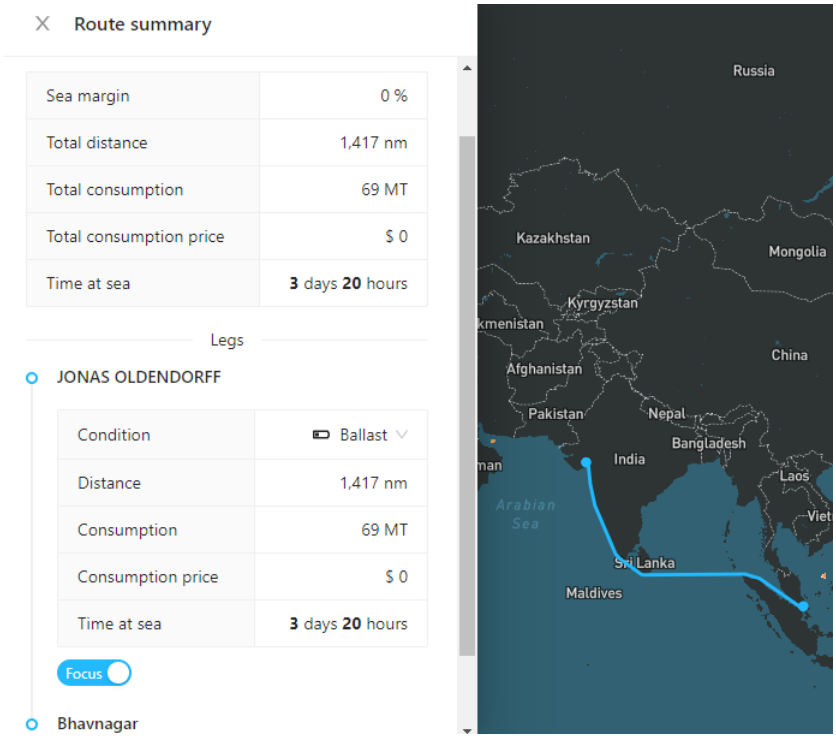


Figure 2.7: Result produced by MO routing engine

data. At present, it's possible to teach computers how to deal with patterns and impart to them a desirable human character. "Analytical Models," known as "Machine Learning Algorithms," allow a computer to learn from data and be trained on it to manipulate processes and make decisions without human intervention. By examining data, models learn to perform a given, generally very specific task. For example, after reviewing thousands of photographs of dogs and cats with labels, the computer will be able to guess if the given image is of a cat or dog.

The general steps to perform machine learning on data is first to train the model where the model will learn the patterns from the data. The followed step is to analyze the trained model and make some predictions on the data which the model hasn't seen it. The final step is to provide feedback to the model so that it can be trained again to get improved results.

Machine Learning (ML) algorithms can be classified into three types which are **Supervised**, **Unsupervised** and **Reinforcement**.

- **Supervised Learning**:
  The ML model is provided by an input variable (x) and an output variable (y) in Supervised Learning [11], and the purpose of the ML method is to learn a mapping function that can learn how the input and output variables

are related. When fresh inputs are presented, supervised Learning aims to anticipate the proper output on the given information that the model has learned from the past samples. Supervised Learning can be further divided into two types of problems which are **classification** and **regression** problem.

  - **Classification**: In classification challenges, ML algorithms attempt to predict which set of classes the given input data belongs to. Depending on number of classes to predict from, classification problems are classified into two types: **binary** classification and **multi-class** classification. In binary classification, there can be only two classes to predict from. Email, for example, falls into the spam or 'nonspam' categories. In multi-class classification, on the other hand, data might belong to numerous classes. For example, to predict a film belongs to which genres 'thriller', 'action, 'romance' or 'comedy.'
  - **Regression**: The primary goal of regression problems is to anticipate the relationship between the dependent (output) and independent (input) variables. In contrast to the classification issue, regression must predict continuous values rather than discrete classifications. Consider stock prices. Because stock prices are continuous, they cannot be chosen as a particular value from a collection of classes.

- **Unsupervised Learning**:
  Unsupervised Learning [12] is a machine learning approach in which the model does not require the users' supervision. This type of learning is beneficial for tasks such as identifying patterns in data that are not explicitly present in the data. As a result, unsupervised learning is often used for data clustering, pattern recognition, and other related kinds of stuff.
- **Reinforcement Learning**:
  Reinforcement Learning (RL) [13] is a type of learning that employs feedback to help a ML system improve its performance. The purpose of RL is to produce the best policy to perform a task based on the system's previous experiences. In other words, RL assists machines in learning how to behave in circumstances where certain behaviors are connected with specific rewards (positive or negative).

Figure 2.8 shows the complete hierarchical structure of machine learning. In this thesis, because there are several ports on which a vessel might arrive, the thesis problem has been characterized as a multi-class classification problem.

2.3 Database

This section discusses some of the concepts of the data which will be used in the thesis. Other table descriptions used in this thesis will be discussed later.

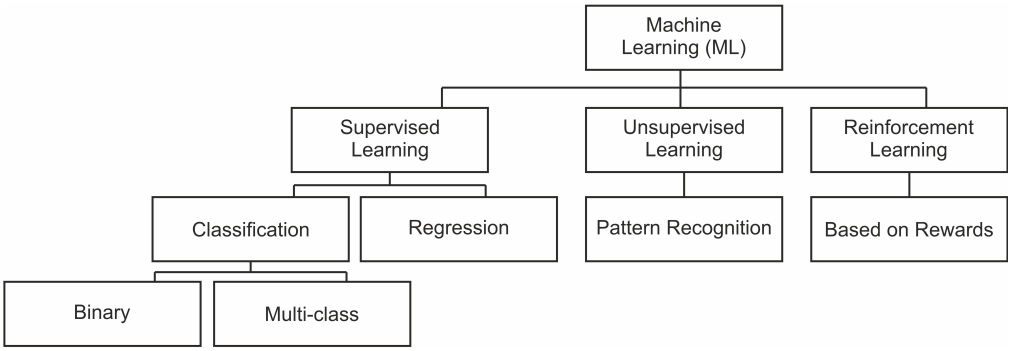


Figure 2.8: Machine Learning (ML) classification

2.3.1 Ports description

The database for the world shipping ports has been taken by the Maritime Optima AS (MO). They have collected all the ports along with their code, name, and geographical location and stored them in a table called 'ports.' But out of all those ports, only some ports are relevant and considered for the trading, so Maritime Optima AS (MO) have marked all the relevant ports visibility as true in the database and for others as false. MO have collected a total of 17365 ports, out of which only 5342 ports have been marked as visible true. The ports have been analyzed and marked visible through the manual process by the shipping experts in MO. The count of visible ports can change according to time. Only visible ports were chosen and utilized for the thesis because shipping professionals indicated that these ports should only be used for prediction. After all, most vessel movements are observed inside these ports.

The UN/LOCODE is used as an identification for the ports. The United Nations (UN) provides and manages this five-letter unique identity of the ports. The first two letters of the five-letter code represent the port's nation of origin, while the following three indicate a more particular location within the country of origin. For example, in locode **NLRTM**, the first two letters define the country 'NL' means the Netherlands and 'RTM' means Rotterdam, which is a city in the Netherlands.

2.3.2 Vessel segments and sub-segments

Maritime Optima AS (MO) have classified the vessels into segments and sub-segments. Segments are defined as the type of cargo usually carried by the vessel. Sub-segments are based on the size, length of the vessels, and also the cargo weight, which is measured in Deadweight Tonnage (DWT) can be carried by the vessel. For example, if a vessel always carries chemicals, then the vessel belongs to the chemical segment. If the vessel is large and carries a large quantity, it belongs to the large sub-segment within the chemical segment. The name for the segments and sub-segment is given by MO's shipping professional, and it follows

the shipping nomenclature. The factors on which the vessels' are categorized into segment and sub-segment have been also defined by the shipping professionals of the MO.

In total, ten segments are been defined by MO to classify the vessels, and they are Dry Bulk, Tanker, Chemical, LPG, LNG, Container, Car Carrier, Oil Service, Combo, Other. For the first eight segments, their name defines what type of cargo they carry. For 'Combo' means vessels that can carry multiple types of cargo. For example, there are vessels that sometimes can carry dry cargo, so they are dry_bulk vessels, or sometimes they carry cars, so they will become car_carrier. And 'Other' includes the vessels like a ferry, passenger vessels, water tankers, and another small kinds of vessels which are usually not of great significant use in the marine trading industry.

Figure 2.9 shows all the segments into which vessels are been divided by the MO experts and also the sub segments are been shown for the LNG vessels.

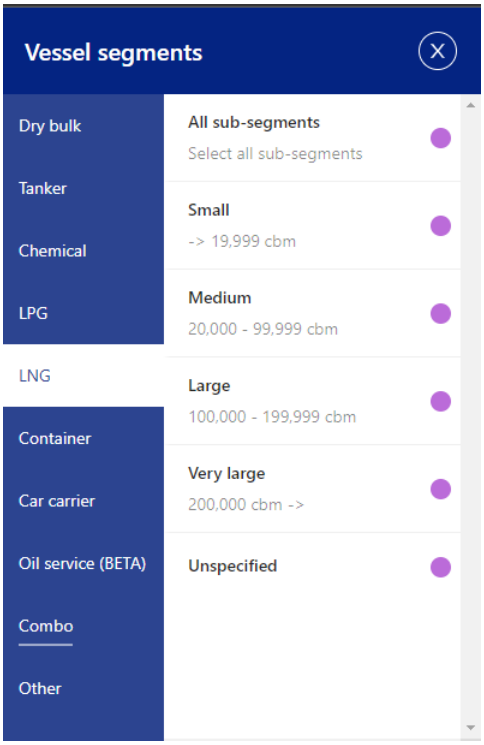


Figure 2.9: All the segments into which vessels are been divided by MO

For the thesis, only five segments have been considered, which are chemical, LPG, LNG, tanker, and dry bulk. These five segments were chosen after speaking with shipping experts in MO, who suggested that these are the five most important vessels in the marine industry, and these five segments carry the most economical

value. Furthermore, the vessels belonging to these five segments typically do not follow the expected routes, making it difficult for humans to predict their destination port. As a result, they indicated that the output of models on these segments would be helpful to know.

2.4 Challenges

This section will focus on initial description of the challenges that were aroused while the development of the thesis solution.

2.4.1 Dataset imbalances

All static and dynamic data which is received by AIS transponders as mentioned in Section 2.1.1 is not always consistent or correct. Static data is more likely to be incorrect because crew members manually input it. In addition, it can be erroneous due to factors such as; crew members forgetting to update the data, inputting the incorrect value, or inputting the value at the incorrect time, resulting in an imbalanced dataset. Many times, dynamic data is also incorrect for a variety of reasons. For example, suppose the signal between the receiver and transmitter is disrupted. In that case, there is a loss of signal for that period, and there is no vessel information for that period, which can range from a few minutes to several hours.

Another issue with AIS data is incorrect mapping of MMSI and IMO values. In the AIVDM/AIVDO protocol, there are primarily two values that are unique to each vessel: the MMSI and IMO numbers. Both of these numbers should be unique for each vessel; however, MMSI numbers can be recycled in some circumstances, such as when a vessel is taken out of service, whereas the IMO number is unique to a vessel's hull. As a result, IMO is the preferable identifier; however, because the AIVDM/AIVDO protocol splits these IDs into positional and static reports, both must be evaluated if static and positional AIS information is to be used.

2.4.2 Machine Learning(ML) challenges

Machine Learning (ML) models are entirely reliant on the data used to train them. The data must be in a specific format and be consistent to get better outcomes. As a result, there are several obstacles when transforming data into a machine learning appropriate format, and their relevant description is covered here.

Categorical column encoding

Categorical columns are defined as columns with finite labels and non-numerical values. On the other hand, numeric columns are columns whose values can be

any valid number. It has been discovered that ML models produce better results when all of the columns are numeric. However, in the data used in the thesis, there are categorical columns; arrival port is a categorical column that has been predicted, along with departure port and other categorical columns that will be added later as a feature in the dataset. So all these categorical columns need to be converted to numerical columns. There are several encoders available for converting a category column to a numerical column, but the most prominent and used ones are **Label Encoder**[10] and **One Hot Encoder**[11]. Label encoder takes a categorical column from data and assigns a number to all of the unique labels in that column; numbers vary from 0 to the column's total number of unique values. It is straightforward to implement; however, the difficulty with this sort of encoder is that ML models attempt to build a pattern between the numbers. As a result, it is not suggested to use it when the rows of data are not connected. On the other hand, in One Hot Encoder, for each categorical column, a new binary feature is created, and the feature of each sample that corresponds to its original category is given a value of 1. However, if a column has a significant number of unique values, the size of the dataset is significantly increased, so when trying to learn from a high number of features, ML models appear to become confused and gets too complicated, as a result, they give bad results. In the thesis, there were many unique values in the categorical column, so that One Hot Encoder will greatly increase the dataset size. On the other hand, if Label Encoder is used, then the numbers will assume a pattern, but there was not any pattern in the data of the thesis. So it was a challenge to decide which encoder should be used.

Data inconsistency

The dataset in which some of the classes account for the majority of the data and the majority of the classes accounts for a few data points is always cumbersome in ML as models tend to see more specific samples than others, making the model partial towards the recurring outcomes. Since few values appear more frequently than the rest during the evaluation process, this may lead to fallacious accuracy values. For instance, take into consideration a binary classification problem having the email as either 'spam' or 'nonspam.' Suppose the dataset consists of 87 samples, where the email is port 'spam,' and 13 emails are 'nonspam.' In this scenario, a simple function can predict an email as 'spam' with 87% accuracy. However, this accuracy would be false since the function will not indicate the email as 'nonspam' irrespective of the input data. In ML models, an identical situation can occur because they are trained on datasets with a disproportionate representation of cases. Some ML models deal with the problem of imbalance better than others, especially decision tree ensemble methods such as the XGBoost model [1]. However, these models might still struggle with highly imbalanced datasets. Visualizing the data that need to be predicted in the latter part of the

[10]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
[11]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

thesis is discovered to be highly inconsistent. Therefore, it was a challenge to make the ML model perform well on the highly discrepant data.

Chapter 3

Related works

A literature evaluation was undertaken to determine the latest advances in the subject area and determine to what degree the literature answered the proposed research questions.

3.1 RQ1.a: What kind models and data have been used to predict the destination of the vessel?

While searching for prior literature on the area of the study, which is the prediction of the port of arrival using AIS data, it was discovered that there are many relevant sources, but with many limitations. The problem with existing studies is that they have been defined or conducted within a limited geographical region. Furthermore, a great deal of research only forecasts future position over short time frames, and other studies concentrate on just finding trajectories for a vessel. Using AIS data, significant research also focuses on collision prevention or identifying abnormalities in shipping patterns or prediction of Estimated Time of Arrival (ETA). According to the subject of interest for this thesis, however, only those papers that concentrate on the prediction of arrival port utilizing AIS data were examined in-depth, regardless of whether they focused on a limited geographical region or the entire world.

The Table 3.1,Table 3.2 below displays published research results about predicting the port of arrival. The tables also shows the features employed by the authors to predict arrival port and the data for which geographic region they have considered in the studies.

Paper	Features	Prediction model	Geo-extent	Accuracy
[14]	Used coordinates from AIS data to focus on transition of vessel on grids	sequence-to-sequence model which uses a spatial grid, specifically LSTM	Mediterranean Sea	1.44 log perplexity
[15]	based on voyage, which is created using positional data, speed and navigational status	Heterogeneous graph based ML model	Region around Danish waters, while predicting ports can be outside	64.72%
[16]	draught, departure_port, trajectory, departure_time, vessel_id	for voyage creation graph based approach and for prediction Recurrent Neural Network	Global for oil tankers	ports: 41%, region: 87.1%
[17]	AIS data for trajectories, distance ratio and distance between two trajectories combined with probability	DBSCAN, Random Forest (RF)	Global	port accuracy: 65.77%, city accuracy: 81.65%
[18]	vessel type, vessel position, speed, course and offset of longitude and latitude from the vessels' positions to all the ports	Neural Network classifier per port	for defined number of ports and vessels	different for different ports
[19]	vessel type, sub-type, AIS positional data	'Venilia' composed of many ML models including Markov models	Large(assumed)	more than 50%
[20]	AIS positional data	nearest neighbour on similar trajectories	Large(assumed)	Accuracy not defined
[21]	speed, longitude, latitude, course, departure port	conventional classification, classification enhanced with clustering, and LSTM based classification; Random forest performs best	Mediterranean Sea	86%
[22]	AIS positional data	Genetic algorithm with some modifications	Two regions in Netherlands	75% for the route extraction
[23]	sheep type, speed, longitude, latitude, course, heading, departure_port, draught	ensemble model based on RF, Gradient Boosting Decision Tree, XGBoost, Extremely Randomized Trees	Mediterranean Sea	97%

Table 3.1: Papers related to the prediction of arrival port (Table 1/2)

Paper	Features	Prediction model	Geo-extent	Accuracy
[24]	ship type, departure port, ship Id, current position	Bayesian Inference and grid-based heuristics	Mediterranean sea	80%
[5]	vessel type, departure_port, similar trajectory destination_port, trajectory_length	Extreme Gradient Boosting (XGBoost)	Global	72%
[25]	AIS positional data along with prior port of the vessel	Random Forest	Baltic sea region	probability is predicted for next ports
[26]	prediction based on trajectory clustering	convolutional auto-encoders combined with clustering and K-means	on a single city area	Not defined
[27]	AIS positional data	genetic algorithm, DBSCAN, directed graph	Large	Not Available

Table 3.2: Papers related to the prediction of arrival port (Table 2/2)

3.1.1 RQ1.a - Summary

There were different kinds of model and features set that have been used to predict the arrival port of the vessels. The papers [5, 17, 18, 20, 21, 23, 25] used the classification approach for the prediction of the arrival port, the papers [15, 16] used the graph-based technique for the prediction of arrival port, while the papers [14, 16, 21] use the sequence to sequence model for their predictions. In most of the papers, the dataset that has been used has been limited to some specific geographical region, which means the number of ports to predict from are also greatly reduced, so the ML models can fit for small data set but for the large dataset, with many classes, the same ML models might not perform well as the inconsistency in the data increases, and also the classes to predict from also increases. It was also observed that the sequence to sequence models are mainly used for the trajectory prediction as the sequence to sequence models are good for the prediction of the next point in time based on time series data [28]. But with the classification approach, different features can be used to predict the arrival port as the arrival port prediction depends on many features like vessel type[5], previous port[25] and further features such as course, heading and draught values [23]. So in the classification model, all these can be combined for the prediction, and it can predict the arrival port at any time in the future. The authors of the paper [21] have tried sequence to sequence as well as the classification and clustering approach for the prediction of arrival port, and the results show that the Random Forest (RF) model gives the best result.

3.2 RQ2: What kind of research methods have been used to predict the availability of vessels at a port for a specific cargo?

There has been very little research done in predicting the availability of vessels at a port for a specific cargo in the past. Not a single publication was discovered that focuses explicitly on this objective. Instead, several publications provide predictions about the Estimated Time of Arrival (ETA) and vessel location after a certain amount of time has passed. The authors of the work [29] have made predictions about the probabilities of vessels being available within a particular region of interest. However, the research presented in the publication does not focus on a specific cargo, nor is it capable of making predictions for any point in the future. In the paper [30] the authors have used the time series data of the number of vessels can be available at an inland port. But the paper neither consider for the specific cargo nor make use of any AIS data, and the data available for this thesis is the AIS data to perform the prediction of availability of vessels. After the research, the papers that concentrate on forecasting the availability of vessels at any moment at a port, especially for a particular cargo, are not discovered in any earlier works.

Chapter 4

Methodology

This section will go through all of the development phases that were completed during the thesis' model development in order to predict the arrival port for the vessels and to anticipate the availability of vessels at a certain port for a specific cargo.

4.1 Approach overview

As can be seen in the section Chapter 3, there is no research done in the area of predicting which vessels will arrive to a port; instead, the majority of the research is done in the area of predicting a vessel's destination port or predicting the estimated time of arrival, given a destination port. In this thesis, a method has been proposed for predicting the availability of vessel at a port. After examining the literature and speaking with shipping professionals, the proposed method will combine two steps, the first step being the prediction of arrival port for all the vessels and second, calculation of the Estimated Time of Arrival (ETA) at the predicted port. Therefore, after these two steps, there will be a table with the predicted port and their ETA at that port along with the probability of the predicted port. So, from this table, query can be raised to find out the count of vessels at any specific port at any time interval. For example, if the chemical cargo has to be picked up from Oslo port after a week, to see which all chemical vessels can arrive at Oslo port, the first step will be to predict the arrival port for all the chemical vessels and then the ETA at that predicted port. In the end a query can be raised to find all the vessels which have arrival port as Oslo and ETA of one week.

The Machine Learning (ML) model has been trained on a collection of features to forecast the arrival port, all of which will be detailed in-depth in the following sections, including steps for the creation of those features and the ML model used for prediction. After the prediction of arrival port,Maritime Optima AS (MO)'s routing engine has been used for the calculation of ETA.

4.2 Initial dataset formation

This section seeks to detail all the data tables used in the thesis, the situations under which data was filtered out of these tables, and why. MO provides the initial data tables. This information is then processed locally on a personal computer, and all the tables are saved locally in a PostgreSQL database server.

4.2.1 Automatic Identification System(AIS) data

Automatic Identification System (AIS) is the data source that is been collected directly by the vessels. All the other data tables and meaningful information have been extracted from the AIS data. MO collects the AIS data from more than 700 AIS satellites every second. These satellites return the data of around 85 000 vessels. MO has been collecting this data since December 2019, and it is still going on. As stated in the section Section 2.1.1, all the AIS comes encoded, and a lot of data is inconsistent. So MO have made their algorithms and services to decode the data and remove all the imbalanced data points to retrieve meaningful information from it. AIS data table contains the following fields:

- **id**: a sequential identifier in the table
- **IMO**: a unique number of the vessel
- **MMSI**: a number given to the vessel
- **position**: geographical coordinates of the given IMO and MMSI number vessel
- **timestamp**: unique UNIX time stamp at which information of the vessel is being recorded

The fields shown above are the fields that are used in the thesis, but the MO's AIS data other fields also which are discussed in the Section 2.1.1. MO have mapped the IMO and MMSI number for the different vessels and stored in the table. So it's easier to fetch the dynamic as well as the static AIS data. For the thesis, AIS data collected from December 2019 to February 1, 2022, was used. All other tables created utilizing AIS data also used data till February 1, 2022.

4.2.2 Ports data

As mentioned in section Section 2.3.1, MO have a database of almost all the shipping ports that exist in the world. But MO have filtered out the ports and kept only the relevant ports which are about 5342 out of 17 365. So, for this thesis also, only relevant ports have been used. All the ports have been stored in the table called 'ports' which has all relevant and irrelevant ports. The attributes of the ports table are:

- **locode**: port's unique identifier by which all ports of the world have been identified as discussed in section Section 2.3.1.

- **name**: this column specifies the name of the port.
- **position**: it is the geographical location of the port.
- **visible**: this field is marked 'f' is the port is irrelevant and 't' if relevant

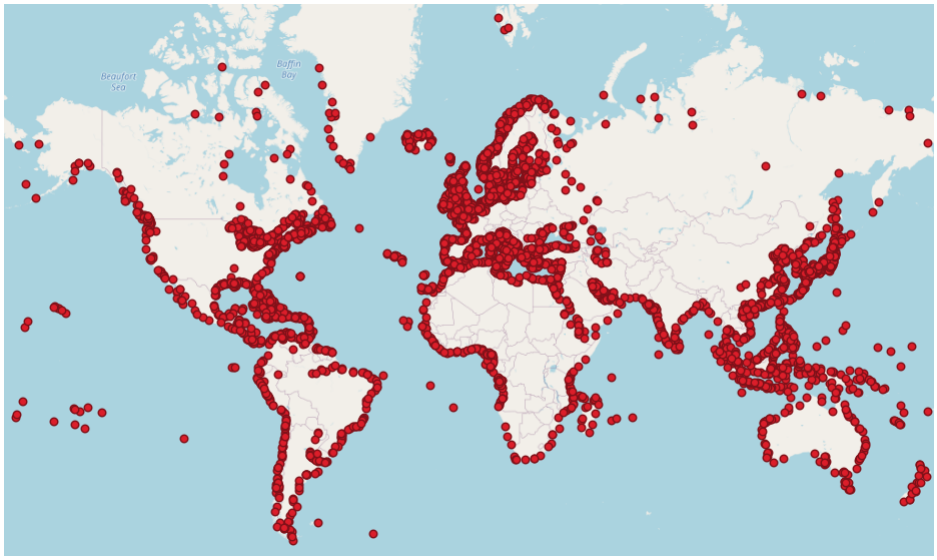


Figure 4.1: Location of all the relevant ports

Figure 4.1 shows the location of all 5342 relevant ports which have been used in the thesis. From the Figure 4.1 it can be analyzed they have been distributed all over the world. But the leading regions are Europe, the USA, and East Asia.

4.2.3 Voyages

For defining the voyage of a vessel as discussed in the section Section 2.2.1, the definition which is provided by MO shipping experts has been used. MO has stored all of the voyages for all the vessels in a table called 'voyages'. This is one of the most important data for the formation of the model in this thesis, as predicting the arrival port is the way of completing a voyage for a vessel that is in the middle, so it is important that ML model sees the solid historical data related to it. For this thesis, the voyages table includes all the voyages, which are based on the AIS data captured between December 2019 till February 1, 2022. The voyage table have the following fields:

- **id**: a unique sequential identifier or it can be called as voyage_id
- **imo**: a unique number of the vessel to which this particular voyage belongs.
- **segment**: defined the segment of a vessel, there can be different segment as defined in Section 2.3.2.
- **sub segment**: it defines the sub segment of the vessel to which it belongs to.

- **departure port**: it defines from which port the vessel have been departed.
- **departure timestamp**: it defines the departure timestamp from the departure port.
- **arrival port**: it defines to which port the vessel has arrived.
- **arrival timestamp**: it defines the arrival timestamp at the arrival port.
- **distance (in meters)**: it defines the distance covered by the vessel from the departure port to the arrival port.
- **duration (in hours)**: it defines the time vessel took to reach to arrival port from departure port.
- **average speed(in knots)**: it defines average speed of the vessel calculated through out the voyage.

This table of voyages has 4395348 voyages, which belong to 61 573 different vessels. However, the 'voyages' table had a large number of inconsistencies. The machine learning model would not have performed well if trained on inconsistent data. As a result, it was necessary to update the voyages table and ensure that the data was consistent for the machine learning model to work efficiently. All of the inconsistencies in data are listed below.

- There were voyages on the table with the same departure and arrival port. Duplicate voyages can occur if the vessel leaves and re-enters the same port geometry or uses the navigational status incorrectly. Maritime Optima AS (MO) filters some of that out but is vulnerable to some miss-use of the navigational status or if there is a significant time delay between leaving and re-entering the same port.
- There were some voyages in which the arrival port or departure port belonged to the ports, which are marked as irrelevant by MO. Historical voyages have been mapped to ports that were deemed relevant at the time, but MO continuously updates this information per port. So, in that case, the port has later been deemed irrelevant, and if MO re-build the voyages, the voyage would have been mapped to a different port if there is another visible one nearby. So MO base their voyages on polygons, radius, and navigational stats as discussed in Section 2.2.1 for visible ports only, but if MO hide a port in the future, they don't update voyage retro-actively, but it will be corrected on future re-builds.
- Some voyages have a distance of 0 while they have different arrival and departure port; practically, this is not possible. This should only be possible if the vessels lie inside a visible port's radius or geometry and switch their navigation status on and off multiple times.

All of the rows which are part of the inconsistency mentioned above have been removed from the 'voyages' table. After deleting all the inconsistent voyages, the final voyage table is left with 2489940 voyages. So these many voyages have been used to build and train the model.

IMO	departure_port	departure_timestamp	arrival_port	arrival_timestamp
5126512	DEHED	2020-10-17 09:12:54	DEWVN	2020-10-19 11:08:26
5126512	DEWVN	2020-10-21 04:22:49	DEREN	2020-10-21 23:48:35
5126512	DEREN	2020-10-22 04:17:07	DEHED	2020-10-22 09:46:55
5126512	DEHED	2020-10-26 08:54:21	DENHO	2020-10-26 22:05:55
5126512	DENHO	2020-10-27 08:45:29	DEHED	2020-10-27 17:35:17

Table 4.1: Voyages for a single vessel in the voyage table

In Table 4.1, it can be seen that the five voyages which have been defined for the vessel whose IMO is 5126512 are all consistent. The vessel is entering and departing ports constantly. Only five segments have been selected for the thesis, as defined in the section Section 2.3.2 which shipping experts have suggested. Therefore the voyages are copied from the 'voyages' table segment-wise to a segment-specific voyage table. So all the LNG voyages have been copied to the 'lng_voyages' table. Therefore all the further calculations on the data have been done separately for all five segments. Table 4.2 specifies the total number of voyages for every segment.

Segment	Total voyages
Chemical	65 238
LNG	15 826
LPG	86 715
Tanker	239 659
Dry_Bulk	997 512

Table 4.2: Total voyages for every segment

4.2.4 Tracks builder

For all the voyages which have been stored in the voyages tables, tracks have been made from the departure port to the arrival port. For the creation of tracks, the positions of the vessel have been retrieved from the AIS data table. The creation of a track will help to find out the route the vessel took to reach the arrival port. All tracks have been stored in voyages tables in a new column named trajectory. The steps followed for the creation of trajectory are listed below:

- **Step 1**: Retrieve all the positions which have been emitted by the vessel during the voyage from the AIS data table between departure_timestamp

and arrival_timestamp sorted by timestamp.

- **Step 2**: In this step, all the positions have been joined by a line to form a trajectory.
- **Step 3**: The last step is to sample the trajectory using the Douglas Peucker algorithm, which has been explained in Section 2.1.3 so it will remove all the extra points from the trajectory while maintaining the shape of the trajectory.

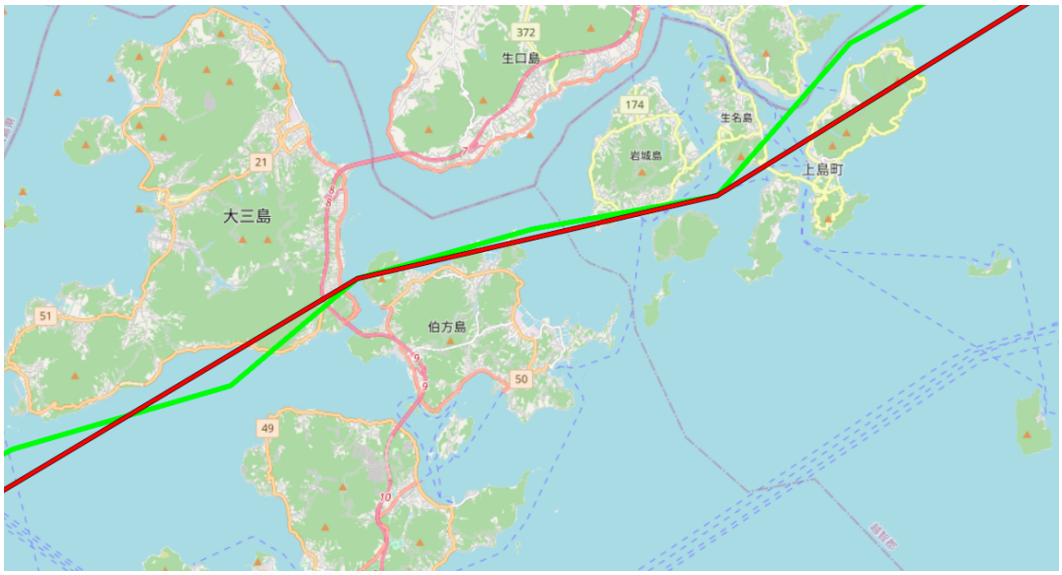


Figure 4.2: Comparison of tracks before and after sampled by the Douglas Peucker algorithm

In Figure 4.2 the 'green' line represents the original trajectory, and the 'red' line represents the simplified trajectory from the Douglas Peucker algorithm. From the Figure 4.2 it can be analyzed that Douglas Peucker has removed many points and created a straight line, especially in starting and at the end of the trajectory.

4.3 Data formation for Machine Learning(ML)

After gathering all initial data along with trajectories, the development of the training dataset that will be utilized to train Machine Learning (ML) models will be the next step. Therefore, in this part, the processes utilized to create the ML dataset will be discussed in depth.

4.3.1 Trajectory Similarity

As discussed in the section Section 2.2.2, vessel arrival port can also be determined by comparing the current trajectory with the past trajectories. This arrival port prediction is solely based on the trajectory comparison; it will not include other

Figure 4.3: Comparison of the current trajectory and the Most Similar Trajectory (MST)

features. The method used to compare the trajectories is Symmetric Segment-Path Distance (SSPD) because this method is proven to give the best results from the past works. The arrival port predicted by this method is called as Most Similar Trajectory's Destination (MSTD). The steps to find the most similar trajectories for the current trajectory are listed below.

- **Step 1**: The first step is to fetch all the past trajectories to compare. All the past trajectories have been fetched, which have departed from the same port as the current trajectory.
- **Step 2**: The second step is to provide the SSPD method with the current trajectory and all past trajectories. This method, after calculation, returns the most similar trajectory along with the distance between the current and most similar trajectory.
- **Step 3**: The destination port of the most similar trajectory, which is returned by the method, is being stored as sspd_mstd and the distance between the trajectories has been stored as sspd_dist in the database.

In the Figure 4.3 the trajectory in the 'red' color is the Most Similar Trajectory (MST) and the trajectory in the 'green' color is the current trajectory. Both the trajectories are leaving from the 'SEFIS' port of Fiskebäck, and reaching the port 'SEMMA,' port of Malmo. This trajectory similarity is shown for the complete tra-

jectory. In a real case, the current voyage will be in the middle of voyage, and the MST will give the destination port for the current trajectory.

4.3.2 Probability

The destination port of the most similar trajectory is stored, as explained in section Section 4.3.1. However, it's also interesting to observe how many other historical trajectories travel to the same destination port after departing from the same port. This probability will indicate chances of mos similar trajectory reaching a MSTD out of all the past trajectories. As a result, it will include the level of trust in the accuracy of the anticipated MSTD. Therefore, it is adding more value to the MSTD which has been predicted by the SSPD method. Therefore, this feature has also been included in the training dataset since it can help predict the vessel's arrival port.

This parameter has been defined as in equation Equation (4.1)

$$probability = \frac{Number\ of\ historical\ trajectories\ with\ MSTD\ as\ destination\ port}{Total\ number\ of\ historical\ trajectories\ with\ same\ departure\_port} \tag{4.1}$$

The code shown in Code listing 4.1 shows the function which will calculate the probability of Most Similar Trajectory's Destination (MSTD) and returns it. The function will receive all the historical trajectories for that voyage, and the selected most similar trajectory according to SSPD method as a parameter. Then the function will loop and count all the trajectories which have a destination port same as the MSTD.

Code listing 4.1: Python code used to calculate probability

```
# This function will receive all the similar trajectory
#  and the most similar trajectory as a parameter.

def get_probablities(cmp_trajs, mst):
    count = 0
    for index,ports in cmp_trajs.items():
        if mst['arrival_port'] == np.array(ports['arrival_port']):
            count = count+1
    probablity = count/len(cmp_trajs)

    return probablity
```

4.3.3 Season

Many voyages are dependent on the seasons, and vessels used to take cargo are based on seasonality. For example, the dry_bulk vessels are used to carry many agriculture-related cargoes. So if the Europe region is considered, there is no agricultural production in the winter season. So, there will be no dry_bulk vessel

voyages in Europe region during the winter season. On the other hand, more energy is required during the winter season in the Europe region, so there will be many LNG and LPG vessels. As a result, this feature has also been discussed with the shipping experts. Their review states that it can be an important feature, and it will be good to keep in the training data set to see its importance while predicting the vessel's destination port.

The seasons after being discussed with experts are grouped by the months as shown in the table Table 4.3.

Season	**Months**
Winter	November - January
Spring	February - April
Summer	May - July
Autumn	August - October

Table 4.3: Seasons grouped by months

The months from departure timestamp have been used to determine the season. It was also proposed that the arrival timestamp can be used to determine the season. However, there will be no arrival time stamp while testing the model on the real data. Therefore, the departure timestamp of the vessel from the departure port is used to calculate the season and assess the feature's relevance for the season.

The seasons are updated in training data directly through query rather than through the code for each voyage. The query used is shown in Code listing 4.2.

Code listing 4.2: SQL Query used to update season in the training dataset

```
create temporary table voyage_seasons as (
        select
                id,
                departure_timestamp,
case
    when departure_timestamp between
        (extract(year from departure_timestamp)::int||'-02-01')::date
        and
        (extract(year from departure_timestamp)::int||'-05-01')::date
    then 'spring'
    when departure_timestamp between
        (extract(year from departure_timestamp)::int||'-05-01')::date
        and
        (extract(year from departure_timestamp)::int||'-08-01')::date
    then 'summer'
    when departure_timestamp between
```

```
        (extract(year from departure_timestamp)::int||'-08-01')::date
        and
        (extract(year from departure_timestamp)::int||'-11-01')::date
    then 'autum'
    else 'winter'
end as season
from chemical_voyages
);

update chemical_voyages set season = t.season
from voyage_seasons t
where t.id = voyages.id;

drop table voyage_seasons;
```

The column named 'season' has already been added to the voyages table to update the season before this query. So the query in Code listing 4.2 picks the departure timestamp to calculate the season based on the timestamp and updates it in the voyages table. In this case, voyages, the table is defined for the chemical segment, so it has been named as chemical_voyages.

4.3.4 Distance ratio

It can be beneficial to see that the vessel is closer to the departure port or closer to the MSTD, which will help to decide the position of the vessel also. Therefore to know this, the distance_ratio has been calculated. The distance_ratio is the haversine distance, which has been defined in section Section 2.1.2, is used. It is one of the best methods to calculate the distance between two geographical points.

In the Figure 4.4 the 'red' trajectory is the trajectory of the vessel. The trajectory depart from the same departure port (*Port B*) and the port predicted by the SSPD method is *Port A*. Distance_ratio is defined as the haversine distance between the vessel's current position (*Cp*) and the MSTD (*Port A*) to the distance between the current position (*Cp*) and the departure port(*Port B*). According to the Figure 4.4 and the equation Equation (4.2), the distance_ratio calculation has been shown below.

$$distance\_ratio = \frac{Haversine(Cp, PortA)}{Haversine(Cp, PortB)} \tag{4.2}$$

, where haversine is the function to calculate the distance between two points as defined in Section 2.1.2. *Port B* is the coordinate of departure port, and *Cp* is the current position; *Port A* is the destination port coordinate of Most Similar Trajectory (MST) trajectory.

So, according to the Equation (4.2) if the ratio is close to or equal to '0', it means the vessel is more close to MSTD, but if the value is substantial, it means that the vessel is still close to departure port. If the value is close to '1', the vessel is mid-way as it is almost equidistant from both the ports. So this feature will help give insight into the position of the vessel; either it is too close to the departure

Figure 4.4: Distance of current position from the departure port and the sspd_mstd

port, which means the voyage has just started, or very close to MSTD, means it has covered a significant distance. It also helps to confirm the value of MSTD. If the vessel is too close to MSTD, then it is most likely to reach the MSTD port only. Therefore this feature has been kept in the dataset.

Code listing 4.3: Python code used to calculate distance ratio

```
def haversine(port,current_position):
    """
    Calculate the great circle distance in kilometers between two points
    on the earth (specified in decimal degrees)
    """
    length = len(current_position)
    lon1 = port[0]
    lat1 = port[1]
    lon2 = current_position[length-1][0]
    lat2 = current_position[length-1][1]
    # convert decimal degrees to radians
    lon1, lat1, lon2, lat2 = map(radians, [lon1, lat1, lon2, lat2])

    # haversine formula
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2
    c = 2 * asin(sqrt(a))
    r = 6371
    # Radius of earth in kilometers.
    #Use 3956 for miles. Determines return value units.
    return c * r
```

The code Code listing 4.3 is the Python code that is used to calculate the hav-

ersine distance. The function accepts two parameters; the first is the port location, and the second is the current position coordinates of the vessel. The end function returns the haversine distance between the port coordinate and the current position. So this function has been called twice, once with the departure port as port coordinate and the second time with the MSTD as port coordinate. And then, both the values have been divided according to the equation Equation (4.2) to calculate the distance ratio.

4.3.5 Creation of training dataset

Some other features have also been added to the dataset. One of them is trajectory_length; this feature will indicate the amount of distance traveled by vessel. Trajectory_length is calculated by counting the total number of points in the trajectory. If the trajectory_length is higher, the vessel is about to arrive at a port, but if it is small, the vessel has just departed. Therefore, trajectory_length has also been kept as a feature. The sub-segment and departure port of the vessel is also being kept as features in the dataset.

As all the voyages in the dataset have been complete, so ML model will not see any incomplete voyages or the vessels which were in the middle or have just started their voyage. Therefore long trajectories have been divided into several small trajectories to create a real scenario to address this problem. This process of dividing trajectories is inspired by the [5], as the author has also done the same thing in the thesis. The author had broken the big trajectory into several small trajectories so that voyages of all the lengths could be seen by the ML models.

After dividing the trajectories into small trajectories, it has been seen that for the small trajectories length (the vessel has just left the port), the MSTD port is predicted wrong. But if the trajectory has a long length, which means the vessel has traveled quite a long distance, the MSTD port has been predicted correctly for most of the cases. So it indicates that for the voyages which have covered a long distance, the MSTD is almost always right. This is sensible since lengthy voyages have long trajectories, making comparisons simple and reducing the number of possible ports. Alternatively, if the trajectory is short, it is difficult to discover a matching trajectory, and there are several port options to choose from.

Table 4.4 shows the snapshot of the voyage, which has been broken down into smaller voyages.

Dividing the trajectories will also help expand the training dataset, as one journey will now be divided into four, and the machine learning model will view the data of the voyages if the vessel is in the middle, beginning, or at the end of the voyage. This process will aid in the training of the ML models and provide better results.

Voyage ID	SSPD-based MSTD	Arrival port	Trajectory length	SSPD dist.
891	JPYOS	JPYSS	3	5935
891	JPCHB	JPYSS	6	4156
891	JPYSS	JPYSS	9	7991
891	JPYSS	JPYSS	12	3042

Table 4.4: A voyage after been divided into smaller voyages

4.3.6 Rejected features

Some more factors are taken into account while creating the dataset. Based on the preceding literature covered in Chapter 3, numerous articles have used the COG, heading, and draught as features for predicting the arrival ports. As a result, these characteristics were also examined for inclusion in the dataset.

The crew members manually input the draught values as specified in the Section 2.1.1. Static data cannot be trusted since it is dependent on the crew entering the information. The crew members intentionally incorrectly input 40% of the data, [31]. Furthermore, according to discussions with maritime data specialists, draught values are mostly recorded towards the journey's conclusion to advise the arrival port about the load. Therefore, its value is mostly zero between the voyages. As a result, in these instances, the draught value is not regarded as a characteristic for prediction.

For the COG and heading value, they have been updated with every AIS message as they belong to the dynamic data. Therefore, these values might alter pretty often over time. Furthermore, dynamic data values might be lost from time to time. As a result, the actual value of the vessel's COG and heading is unknown. As a result, COG and heading are not included as features in this thesis. The COG and heading values may be attempted in the future for testing, but the draught values are typically incorrect and cannot be trusted; thus it is not suggested to be included as a feature for the prediction purposes.

4.3.7 Pipeline for creation of training dataset

As it has been previously stated that the initial data tables had been provided by the Maritime Optima AS (MO), which are the AIS data table, voyages table, and ports table, for all the voyages tracks have been made and stored in the voyages table itself. From the voyages table, segment-wise voyages have been extracted and stored in another voyages table, identified by segment type *_voyages*, where a segment can be any as described in Section 2.3.2. From this segment-specific voyages table, training data set features have been calcu-

lated, which are sspd_dist, sspd_mstd, season, distance_ratio, probability, and the trajectory_length. All of these values have been calculated and stored into a new table called *<segement>_training_data* along with the sub-segment, departure port referenced from the voyage table of that segment. For example, if the training data for the chemical segment have to be created, then the voyages will be fetched from *chemical_voyages* table, and after calculating values for all features, the values will be stored in *chemical_training_data* table. The steps are shown below for the creation of the training dataset

- **Step 1**: In the first step, as discussed in Section 4.3.3 seasons are added through the query Code listing 4.2, so on the voyages table, the query for the season is being run and seasons are being added.
- **Step 2**: In step 2, all the voyages are being fetched through the Python code, and the SQL query for fetching all the details is listed in Code listing 4.4.
- **Step 3**: In the next step, all the voyages are passed and divided into small voyages based on trajectory_length.
- **Step 4**: All the voyages which have been created after Step 3 passed into the function, which will find the sspd_mstd along with the sspd_dist. In this function itself, probability and distance_ratio are calculated. So this function finally returns the value of sspd_mstd, sspd_dist, probability and distance_ratio.
- **Step 5**: After calculation of all the features, the data has been inserted into the training data table.

Code listing 4.4: SQL Query used to fetch the voyages

```
SELECT
          id,
          departure_port,
          departure_lon,
          departure_lat,
          season,
          arrival_port,
          arrival_lon,
          arrival_lat,
          st_x(points) AS lon,
          st_y(points) AS lat,
          sub_segment,
          imo
      FROM (
          SELECT
              -- voyage info
              a.id as id,
              a.imo as imo,
              (st_dumppoints (a.trajectory)).geom AS points,

              -- departure port position and id
              st_x(b.position) as departure_lon,
              st_y(b.position) as departure_lat,
              b.locode as departure_port,
```

```
            -- arrival port position and id
            st_x(c.position) as arrival_lon,
            st_y(c.position) as arrival_lat,
            c.locode as arrival_port,

            a.sub_segment,
            a.season

        FROM "lpg_voyages" a

        LEFT JOIN ports as b ON (b.locode = a.departure_port)
        LEFT JOIN ports as c ON (c.locode = a.arrival_port)

    ) f
    ORDER BY id ASC
```

In table Table 4.5 the final columns of the training data set which have been used for the training can be seen. From the Table 4.5 id, IMO, voyage_id, and mstd_id have been there just for reference. They have been dropped while training the model.

Column	Type	Description
id	serial number	unique identifier
voyage_id	number	the original voyage id from voyages
imo	text	identifier for the traveling vessel
mstd_id	number	identifier of the most similar trajectory's destination
segment	text	the vessel's segment
sub_segment	text	the vessel's sub-segment
departure_port	text	UN/LOCODE of the vessel's departure port
trajectory_length	number	number of points in the trajectory
sspd_mstd	text	UN/LOCODE of the MSTD value for the voyage trajectory
sspd_dist	number	a measure of how similar the voyage's trajectory is to the most similar historical trajectory
probability	number	as described in section Section 4.3.2
distance_ratio	number	as described in section Section 4.3.4
season	text	season based on departure timestamp
arrival_port	string	UN/LOCODE of the vessel's arrival port

Table 4.5: Final structure of the ml_training_data database table.

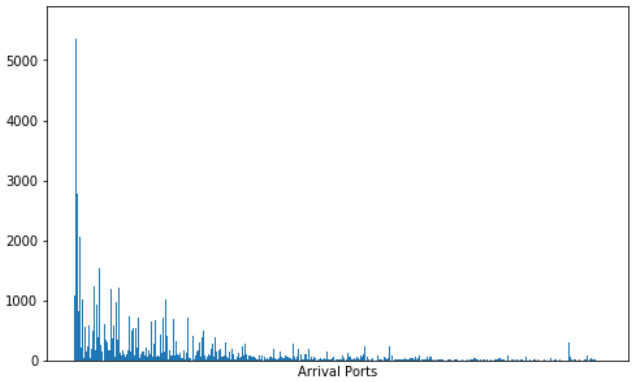
4.4 Machine Learning(ML) experiments

Following the creation of the Machine Learning (ML) training dataset, the next step is to select a suitable ML model, tune it, and train it to predict values for the arrival port column in the training dataset. All ML experiments are run on chemical segment training data, and then the chosen model with the chosen hyperparameters is trained and evaluated on other segments. The chemical segment has been chosen because it has a fair number of voyages, neither too few nor too many. If there are too many journeys, experiments will require a large amount of time to execute and test the results; if there are too few, the model may fit in this segment but not in others. For chemical vessels, according to shipping experts, it is difficult to predict patterns also because they do not sail on the same routes. As a result, if the model fits the chemical segment, it is likely to fit other segments as well. From the selection of the ML model to the final prediction of the arrival port using the ML model, this section describes all the steps in detail.

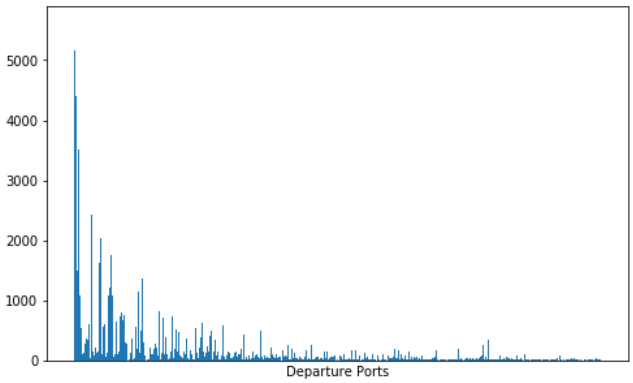
4.4.1 Visualizing data

Before creating a ML model is always good to visualize the data on which the model is going to be trained. As ML models run on the data and tries to analyze the patterns in data, it's important that data is consistent so that the model understands the pattern between the data quite well and returns the best result. Due to this reason, as the first step, data is being analyzed and made consistent before training the model. Some of the statistics of the data of chemical segment are as follows:

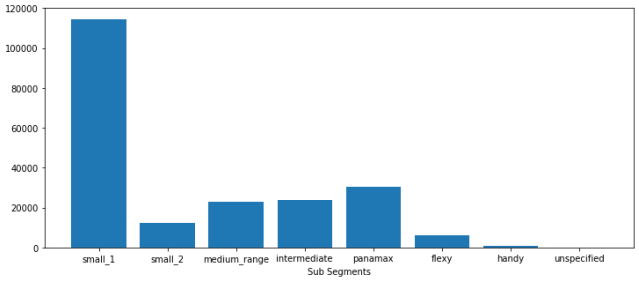
- There were 1503 different arrival ports in the dataset. It can be analyzed from the Figure 4.5a arrival ports distribution is totally inconsistent.
- There were 1387 different departure ports in the database. It can be analyzed the Figure 4.5b departure ports distribution is also totally inconsistent.
- There were a total of 8 different sub-segments in the dataset. It can be seen from the Figure 4.5c that most of the data points are for the small subsegment and there are only some of the data for the large. But not too much inconsistency.
- There are four seasons. It can be analyzed from the Figure 4.5d that the distribution for the season is consistent.
- There were a total of 18058 unique voyages in the dataset. Unique voyages mean the unique combinations of arrival and departure ports the vessel has traveled.



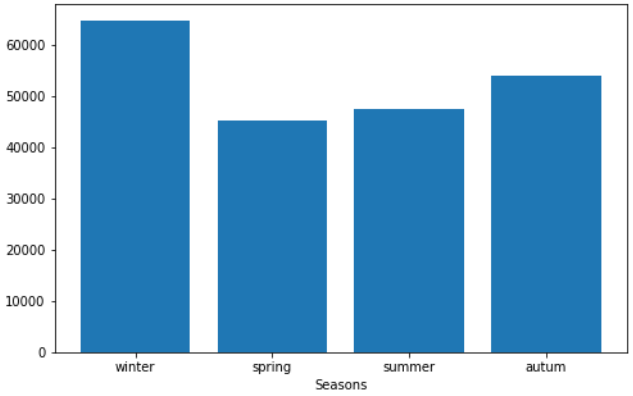
(a) Distribution of Arrival Ports



(b) Distribution of Departure Ports



(c) Distribution of Sub-Segments



(d) Distribution of Seasons

Figure 4.5: Distribution of different categorical features across training dataset

4.4.2 Data preprocessing

Data Consistency

From the aforementioned Section 4.4.1, it can be inferred that the data is very inconsistent. From the Figure 4.5, it can be seen that arrival port data is skewed, only a few arrival ports account for most of the data, and a large number of ports account for very few data points. Therefore, to make the data consistent for the ML model to perform well, the data have been made consistent.

Therefore to make the data consistent, the arrival ports which occur for few times have been deleted. Because these ports will be seen very few times by the ML model, the pattern to predict these arrival ports will not be learned by the model. Therefore, the model will not predict these ports, regardless of whether it has been trained on the entire dataset. Furthermore, the model will become quite complicated by getting trained on the whole dataset. With so many distinct classes to forecast, the outcomes will also worsen for the arrival ports that occur the most often. Because of this, arrival ports that have been used only a few times were eliminated. According to the shipping experts, it is acceptable to eliminate arrival ports from the dataset if they appear only a small number of times. Their opinions are described in further detail under the Section 5.5.

As the first step, the departure port and arrival port combination was seen to remove the arrival ports. Because by looking at the combination, the whole voyage has been looked at. If the vessel appears to take a particular voyage less number of times, it's good to remove the data points for that voyage. There were a total of 18058 voyages in the chemical dataset. Only 12000 voyages accounted for more than 90% of arrival ports. So 66% of the voyages consisted of more than 90% of the arrival ports. Therefore, all the arrival ports in the remaining 34% of the voyages have been removed from the dataset.

In the Code listing 4.5 from *Command 1*, it can be seen that 12000 values accounts for more then 90% of the arrival ports. *Command 2* shows that arrival ports that have combinations of arrival and departure ports occurring four or less than four times can be removed. *Command 3* removed all the data points where the combination of arrival and departure port was less than four times. The graph of arrival ports after removal of voyages is shown in Figure 4.6

Code listing 4.5: Python code to remove inconsistent data based on combination of arrival and departure port

```
# Shows that 12000 values accounts for 90% of the data
Command 1:
df.groupby(['departure_port','arrival_port']).size().sort_values(ascending=False)
[:12000].sum()/len(df)
#Output:
0.9044381545697204
```

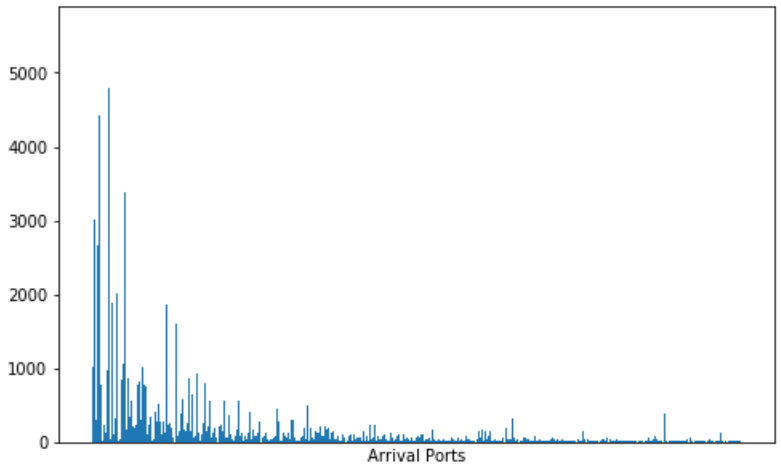


Figure 4.6: Arrival ports distribution after removing voyages appearing less than 4 times

```
# This command shows the count for the combination of arrival and departure port
# till 12000
Command 2:
df.groupby(['departure_port','arrival_port']).size().sort_values(ascending=False)
[:12000]
#Output:
departure_port  arrival_port
JPTND           JPETA          1224
JPETA           JPTND           931
JPCHB           JPNGO           522
JPNGO           JPCHB           518
JPSKD           JPYKK           498
                               ...
EENAI           EESLM             4
                FITOK             4
EEMUG           RUKDT             4
Length: 12000, dtype: int64

# This command will filter the arrival ports based on the given condition.
Command 3:
df2 = df.groupby(['departure_port','arrival_port']).filter(lambda x: len(x) > 4)
```

After removing the values mentioned earlier, the unique values of the arrival port were reduced to 962 from 1503. After further research, it was seen from Figure 4.6 that the arrival ports were still skewed and could be reduced further. Now only the arrival ports have been seen. It was found that only 400 ports occur more than 92% of the time. So 41% of the ports appeared 92% of the time. Therefore, all the other 59% of the ports were removed from the dataset.

Code listing 4.6, *Command 1* showed that only 400 ports out of 962 ports accounted for almost 92% of the data. *Command 2* showed that the occurrence of arrival ports less than 59 times should not be part of the dataset. *Command 3*

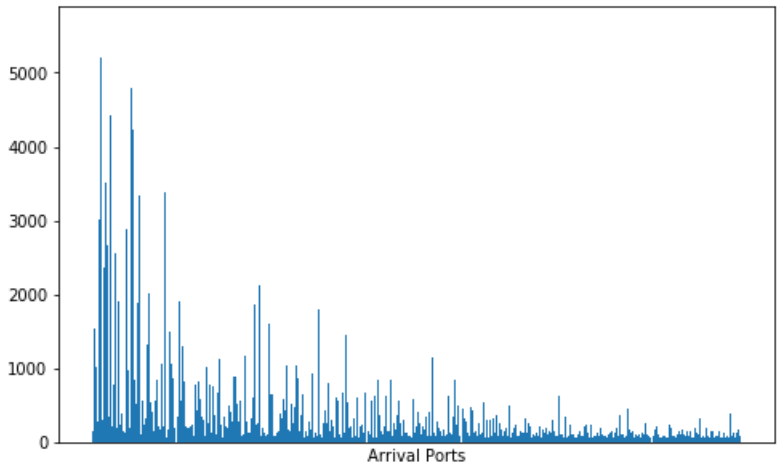


Figure 4.7: Arrival ports distribution after removing voyages appearing less than 4 times and the ports which occurs less than 59 times

keeps only ports that occur more than 59 times in the database.

Code listing 4.6: Python code to remove arrival ports which accounts for less data

```
# Shows that 400 values accounts for 92% of the data
Command 1:
df2.arrival_port.value_counts()[:400].values.sum()/len(df2)
#Output:
0.9279144235311959

# This command shows the arrival port counts for the first 400 ports.
Command 2:
df2.arrival_port.value_counts()[:400]
#Output:
port     count
JPNGO    5204
JPYKK    4801
JPMIZ    4415
JPSAK    4227
JPCHB    3521
         ...
AOPSA      60
GUAPR      60
TWTPE      59
Name: arrival_port, Length: 400, dtype: int64

# This command will remove all the ports which occur less than 59 times.
Command 3:
removals = df2['arrival_port'].value_counts().reset_index()
removals = removals[removals['arrival_port'] > 59]['index'].values
removals.size
df3 = df2[df2['arrival_port'].isin(removals)]
df3.arrival_port.value_counts()
```

After this, the graph of the arrival port is shown in the Figure 4.7. It's still

skewed, but not as much as it was before and number of classes are also reduced from 1503 to 399. So, the machine learning models on this dataset can be run.

Categorical column encoding

As mentioned in section Section 2.4.2, for the ML models it is essential to convert the categorical values into numerical values. There are two encoders mentioned in Section 2.4.2. Both of the encoders have been tried, but with One-Hot Encoder, the size of the dataset has increased drastically. There were around 500 unique departure ports, so new columns have been created for all departure ports. Similarly, there are around 800 unique sspd_mstd values, so new 800 columns have been added, and some other columns are being added for segment and season values. So this makes it practically impossible for the ML models to analyze the pattern out of so many values. Therefore, Label Encoder has been used in this thesis to encode all the categorical values of the dataset.

Scaling of numerical features

When applying machine learning algorithms to a data set, scaling the numerical data is also an essential data preprocessing step. Suppose the numerical data in any circumstance has data points that are far apart. In that case, scaling is a strategy for bringing them closer together, or, to put it another way, scaling is used to make data points more generic so that the space between them is reduced. The model's results are more imprecise when there are more significant differences between the data points of input variables. Machine learning models provide weights to input variables based on their data points and output inferences. In such a situation, if the difference between the data points is substantial, the model will need to give the points more weight, and the model with an enormous weight value is generally unstable in the end. This implies that the model may give unsatisfactory results or perform badly during training.

So for the reasons mentioned above, sampling is done for all the numerical columns: probability, distance_ratio, trajectory_length, and sspd_dist. For the sampling purpose *MinMaxScaler* is been used.

MinMaxScaler divides by the range after subtracting the feature's minimal value. The range is the difference between the maximum and least values at the beginning. MinMaxScaler preserves the shape of the original distribution. MinMaxScaler returns a feature with a default range of 0 to 1.

The final output training data after all the preprocessing can be seen in the table Table 4.6. All the numerical column values are between 0 and 1, and all the categorical columns are changed to numerical values.

sub_segment	departure_port	trajectory_length	sspd_mstd	probability	distance_ratio	sspd_dist	season
1	1	0.000252	1	0.001278	0.370006	0.000812	1
5	1	0.001007	1	0.001221	0.370006	0.002160	1
1	1	0.001761	2	0.000156	0.518240	0.000049	2
1	2	0.000001	3	0.000568	0.427642	0.000119	2
2	2	0.000503	3	0.000490	0.427642	0.000041	3

Table 4.6: Machine Learning Data after PreProcessing

4.4.3 Model selection

After the completion of the data preprocessing steps, the data is ready to be trained by the ML models. Since the prediction of an arrival port is to be done out of 400 different classes, this problem falls under the multi-class classification problem. So for this thesis, several classification models have been tried out that support multi-class classification. After the preprocessing step, all the models have been tried out on the final data. Finally, all the tried models, along with their accuracy, have been presented in the table Table 4.7.

Model	**Acc. %**
XGBoost	73.97%
Random Forest	71.9%
Keras DNN classifier	66.5%
k-Nearest Neighboor	57.4%
Naive Bayes	51.55%

Table 4.7: All tried models along with the accuracy's

From the Table 4.7, it can be observed that K-Nearest Neighbour and Naive Bayes perform worst, so they have been removed in the first step without being researched further. For the Deep Neural Network classifier improvement has been made by tuning the hyperparameters. Many layers have been added, several loss functions have been used, and several optimizers have been tried. As an infinite number of possibilities can be tested to improve the accuracy of the deep learning model, this model has also been dropped as, after several tries, it fails to give accuracy higher than XGBoost or RF models.

After many model trials, it was inferred that the tree-based classification model performed the best when the data was skewed, and there were many classes to predict from. There are two famous tree-based models which are generally used for the classification: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Both of these models seem to give similar results, as can be seen from the Table 4.7. But as the Extreme Gradient Boosting (XGBoost) model's accuracy is a bit higher, consumes less memory, and supports both out-of-core and incremental learning, the implementation is also more efficient in handling bigger data volumes. Therefore Extreme Gradient Boosting (XGBoost) model has been selected as the final model for predicting the arrival port of the vessels.

4.4.4 Training process

For the training of ML models, the dataset had been split between the train and test datasets. The split was based on an 80% - 20% ratio, which means 80% of

the dataset is used for training and 20% for validation. The accuracy which is presented in Table 4.7 is the accuracy that the model gives on the validation dataset, as this is the unseen data by the model, so it is always best to assess the model.

There are many hyperparameters involved in the Extreme Gradient Boosting (XGBoost) model. Hyperparameters are the parameters that control the learning process of the model. A model cannot estimate or set hyperparameters by itself; they have to be set externally. One of the most important steps is to find the best combination of hyperparameters to be set in the model to get the best results. The hyperparameters of Extreme Gradient Boosting (XGBoost) model are discussed below:

- *max_depth*: it defines the maximum depth of each tree in the model. A deeper tree may improve performance, but it also adds complexity and the risk of overfitting. The default value is 6.
- *subsample*: states the percentage of samples from the training data that have been used to create each tree. The default value is 1.0.
- *colsample_bytree*: it is defined as a number of features to be used while constructing a tree. It can help in improving overfitting. Lower values avoid overfitting, but they may also result in underfitting. The default value is 1.0.
- *min_child_weight*: the minimum weight required for a child node. The default value is 1.0.
- *gamma*: the smallest loss reduction necessary to separate a node in a tree. The default value is 0.
- *learning_rate*: the learning rate defines the step size at each iteration. A low learning rate slows computation and necessitates more rounds to accomplish the same residual error reduction as a model with a higher learning rate. However, it maximizes the chances of achieving the best possible result. The default value is 0.3.

For finding the best hyperparameters, there are two methods RandomizedSearch and GridSearch. A Random Search takes a vast (potentially infinite) range of hyperparameter values and iterates over them randomly for a defined number of times to find the best possible combination of hyperparameters. GridSearch tries all the combinations of the given possible values (a finite number for each hyperparameter) and returns the best combination possible out of all the given values. RandomizedSearch method is used on a finite number of values to find the best hyperparameters, but as GridSearch, it will not try all the combinations. It will pick up random values and return the best possible combination of hyperparameters. In Code listing 4.7, it can be seen that a defined number of values have been given for all the parameters, and the RandomSearch function will randomly iterate over these values and gives the best parameters values which can be used for training the model.

Code listing 4.7: Python example showing RandomSearch function to find best hyper parameters

```
from sklearn.model_selection import RandomizedSearchCV
import xgboost
classifier = xgboost.XGBClassifier()
params = {
 "learning_rate" : [0.05,0.10,0.15,0.20,0.25,0.30],
 "max_depth" : [ 3, 4, 5, 6, 8, 10],
 "min_child_weight" : [ 1, 3, 5, 7 ],
 "gamma": [ 0.0, 0.1, 0.2 , 0.3, 0.4 ],
 "colsample_bytree" : [ 0.3, 0.4, 0.5 , 0.7 ]
}
rs_model=RandomizedSearchCV(classifier,
param_distributions=params,
n_iter=5,n_jobs=-1,cv=5,verbose=True)
rs_model.fit(X_train,y_train,eval_metric = ['merror','mlogloss'])
return rs_model.best_params_
```

Finally, on the best hyperparameter values, the model has been trained. Two evaluation metrics have been used to measure the performance of the model on the training datasets: *logarithmic loss* and *classification error*. Both of these are the best and most commonly used metrics to measure the performance of the Extreme Gradient Boosting (XGBoost) model during training; therefore, they have been used. If values for both of these metrics are reducing during the training process, then the Extreme Gradient Boosting (XGBoost) model is learning and improving with every step. Early stopping is also used to ensure that the model is not overfitting on the data. Code listing 4.8 shows the training of the model and the value of hyperparameters that have been used.

Code listing 4.8: Python code showing the training of XGBoost model

```
clf = xgb.XGBClassifier(objective='multi:softmax',seed=42,
learn_rate=0.1,max_depth=6,gamma=0.29,
subsample=0.9, colsample_bytree=0.5)
clf.fit(X_train,
        y_train,
        verbose=True,
        eval_metric = ['merror','mlogloss'],
        early_stopping_rounds=12,
        eval_set=[(X_test,y_test), (X_train, y_train)])
```

After the model is trained on the training dataset, it is evaluated on the test dataset, and the accuracy is calculated. Code listing 4.9 shows the code for calculating the accuracy of the test dataset.

Code listing 4.9: Python code to calculate the accuracy value

```
y_pred = clf.predict(X_test) # model will predict the values of test dataset

# this will calculate the accuracy score and return the accuracy percentage value
predictions = [round(value) for value in y_pred]

accuracy = accuracy_score(y_test, predictions)

print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

The results for all the segments have been presented and discussed in detail in Chapter 5.

4.4.5 Pipeline for running Machine Learning(ML) models

After performing many experiments and data processing steps, Extreme Gradient Boosting (XGBoost) model with the selection of best hyperparameters has been finalized. So for the training of every segment vessels, below mentioned pipeline has been followed:

- **Step 1**: Visualize the data and make the data consistent following the steps mentioned in Section 4.4.2.
- **Step 2**: After making the data consistent, apply the Label Encoder and MinMaxScaler to make the dataset values compatible for the model training.
- **Step 3**: Split the dataset into training and test data.
- **Step 4**: Train the Extreme Gradient Boosting (XGBoost) model on the training data using the same combination of hyperparameters.
- **Step 5**: Evaluate the accuracy of the model on test data.

It is important to save and download the trained ML model so that it can be loaded for predicting the arrival port in the future. In addition, label Encoder and MinMaxScaler have also been downloaded for every column separately. This is because they need to encode future data values in the same way as they are encoded while training.

4.5 Prediction for the availability of vessel at a port

As it was mentioned before, to predict the availability of the vessel at the port, there were two steps. The first was to predict the arrival, and the second step was the calculation of ETA for that predicted port. For the first step, the ML model is run to predict the arrival port for all the vessels of a segment that are in the middle of their voyage. In the next step, the ETA will be calculated to the predicted arrival port for all vessels. In this section, all steps performed to predict and validate the availability of vessels at a port will be discussed.

4.5.1 Data creation

For the first part, that is, the prediction of arrival port, ML model has already been defined. So, now it can be used to predict the arrival port for all the vessels which are in the middle of their voyages.

One date has been selected to find all the vessels which are currently in the middle of the voyage, and all the vessels that start their voyages before that selected date but end after the selected date have been found. So the model predicts

arrival ports for all those vessels at that selected date. The selected date which is chosen is February 1, 2022, because the ML model has been trained for all the voyages which have been recorded till February 1, 2022, only. So all the arrival ports have now been predicted for the voyages, which are never seen by the trained ML model. In this way, the ML model will also be validated on the completely unseen data.

In Code listing 4.10, using the query, from the voyages database all vessels whose voyages began before February 1, 2022, but concluded after that will be retrieved . In addition, the query will also retrieve all information required for the construction of all feature sets according to the training data regarding a vessel from the voyages database. The processes which are undertaken to prepare the data to be equivalent to the training data so that the arrival port may be predicted based on this data are listed below:

- **Step 1**: For all the extracted vessels' voyages, make the track using the track builder, which is explained in Section 4.2.4.
- **Step 2**: Pass the data of voyages with the tracks to the training set builder pipeline mentioned in Section 4.3.7, which will return the data with all the features required to predict the arrival port from the ML model.
- **Step 3**: Insert all the data in a new table, which has been defined segment-wise as *_test_data*. So, for the LPG voyages the test data will be saved in *lpg_test_data*. It is named 'test_data' because all these voyages have not been seen by the ML model. So one of the purposes is also to test the accuracy of the ML model on the real data.

Code listing 4.10: Python code showing the extraction of vessels that are in middle of their voyages

```
def get_vessels(connection):
    q = """
        select departure_port, id, imo,
        departure_timestamp, sub_segment,
        arrival_port
        from lpg_voyages
        where '2022-02-01' between departure_timestamp and arrival_timestamp;
    """
    try:
        result = select(connection, q)
    except Exception as e:
        raise e

    return result
```

4.5.2 Arrival port prediction

After the preparation of data for which the arrival port must be predicted, the next stage is to provide the data to the ML model for the prediction. However, before sending data, categorical columns must be transformed into numerical columns

using the same encoders that were used during the training phase and are also being downloaded in the end so they can be used in future predictions. The same encoders have to be used because, if the Label Encoder assigned the value '200' to the departure port 'NOOSL' during the training, then it should assign the same value to the new data as well. If a different value is assigned to '200', then the model should consider '200' as the new value, but it will treat '200' as 'NOOSL' because it has been trained on it, which will lead to incorrect predictions.

For using the same encoders as the training data, it also needs to be made sure that there are the same values as in the training data categorical columns. But as many data points have been removed while training to make the data consistent, it might not be possible that the Label Encoders will encode all the departure port and sspd_mstd values of the test_data. The seasons and sub-segments have only a few classes, so after removing most of the data also, all the classes of season and sub-segment will be maintained in the data set during the training. Therefore, from the test_data, all the departure ports and sspd_mstd on which the model has not been trained have been removed. While removing all the departure ports and sspd_mstd, it was seen that there were very few voyages for which the departure port and sspd_mstd model had not been trained. Exact figures had been provided in Table 5.5, so it was reasonable to remove the data for the arrival ports that occur very few times.

After removing ports, all the data has been encoded using the specific encoders for the specific columns. In the end, the final data has been passed to the ML model for the prediction of the arrival port. The method called *pred_proba()* of Extreme Gradient Boosting (XGBoost) has been used to predict the port. This method will return the probability of all the classes which are predicted by the model. Among all the classes, the class with the highest probability is selected as the arrival port for that vessel's voyage. So probability can indicate the chances of reaching the arrival port. If the probability is higher, the ML model is confident that the vessel will reach the predicted port. If it's lower than the predicted port, it might be wrong. For the remaining vessel's voyages which are not being predicted by ML models, they have been given predicted arrival port as the sspd_mstd arrival port, and the probability is also given the same as the Section 4.3.2, which has been calculated during the preparation of training data set.

So the final predicted ports are the combination of predictions by ML models and the trajectory similarity method. Its also been seen that SSPD is a good way of trajectory similarity, and for long trajectories, most of the time, it returns the correct ports. In the end, the final predicted ports have been compared with the actual arrival ports, and the results have been presented in the Section 5.3.1.

4.6 Calculating Estimated Time of Arrival(ETA)

After the prediction of the arrival port, the last step is to calculate the arrival time to the predicted port. For the calculation of arrival time, Maritime Optima AS (MO) shipping expert suggested using the Routing Engine, which is developed in MO. This gives the best optimal time to the port and checks all the constraints like if the vessel is big, it will not go through the Suez Canal or Panama canal and avoid the land and ice if there is any in the route. So it is one of the best methods to calculate the arrival time.

MO's routing engine take the arrival port coordinates and the vessel's current position coordinates, as well as the speed of the vessel. The speed at the current position of the vessel has been used for the calculation of ETA. The routing engine calculates the best possible time and returns the time (in hours) that the vessel takes to reach the predicted port. For the thesis, the current position of the vessel is the position which has been reported by the vessel on February 1, 2022, as all the arrival have been predicted considering February 1, 2022, as the present date.

After the calculation of the estimated time of arrival, final table has been made, which consists of IMO, departure_port, predicted_arrival_port, predicted_probability, and ETA which is in hours. This table has been saved for all the segments separately. Now, if the prediction is to be made for the number of vessels that can show up at Oslo port after one week to pick up chemical cargo, then use the final table for the chemical segment. Query all the vessels which have predicted port Oslo and the ETA to be equal to or more than 96 hours. It will return all the vessels with chances of arrival. In the Table 4.8, data that is stored in the final table of the chemical segment can be seen.

imo	departure_port	predicted_arrival_port	predicted_probability	ETA(in hours)
9439785	MXTPB	USRCH	66.66%	810.66
9829409	ESHUV	JPYKK	80.81%	2100
9323338	MYSUP	INNML	33.33 %	123.70
9288930	CNDAL	CNTNG	13.10%	88.22
9380386	BRMAO	USNXL	56.33%	51.85

Table 4.8: Final table

4.7 Summary

The summary for the prediction of the availability of vessels at the port has been explained here with an example to be relatable to the real-world scenario. Let's take an example from the broker's perspective if the cargo owner comes to the

broker with an urgent need to transport the LPG from the Oslo port to the port in China. Now port owners have to find out the earliest possible arrival of Liquefied Petroleum Gas (LPG) vessels at the Oslo port. So to predict this, the steps that broker have to perform are as follows:

- **Step 1**: Find all the LPG vessels which are in the middle of their voyage.
- **Step 2**: Predict the arrival port for all those vessels using both ML model and trajectory similarity method.
- **Step 3**: Calculate the Estimated Time of Arrival (ETA) at the predicted arrival port.
- **Step 4**: Search for the earliest availability of vessel at Oslo port from the final table.

The Figure 4.8 shows the overview of the steps to perform to predict the availability of vessels at a port. In the end, on the 'Final Table' query can be run to get the desired results.

4.8 Methodology conclusion

All of the steps performed in the thesis have been thoroughly detailed throughout this chapter. The following chapter will go through the outcomes and validation of all the procedures explained in this chapter.

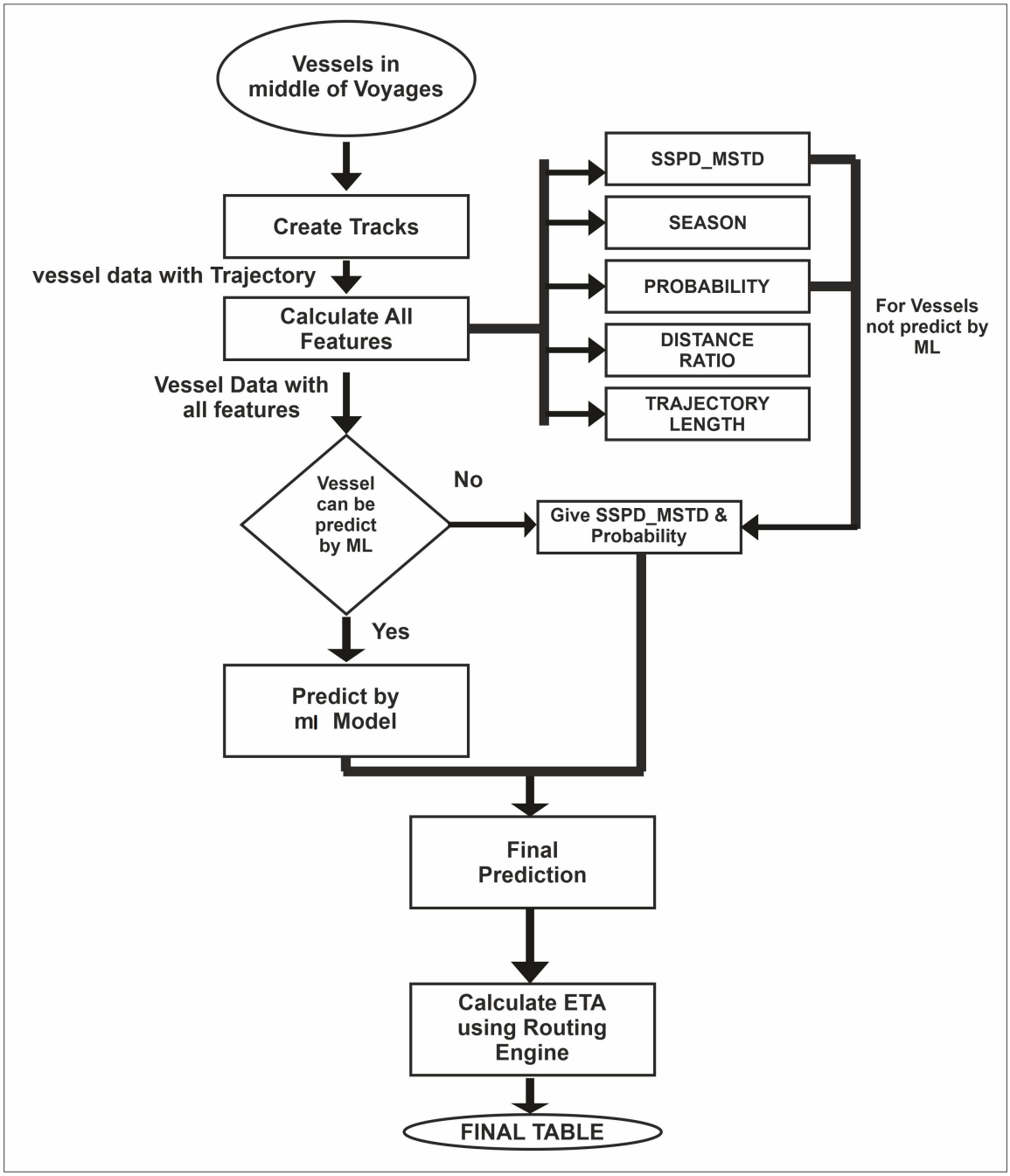


Figure 4.8: Flow Chart for the prediction of availability of vessels at a port

Chapter 5

Results

This chapter includes the results of the proposed solution and the changes performed based on the analysis of the results. Furthermore, this section will have the reviews of the experts from the shipping industry on the solution, so it will help determine the validity of the solution.

5.1 Dataset validation

The initial dataset was supplied by Maritime Optima AS (MO) which consists of the voyages table, the ports table, and the AIS data table. From the dataset, the data for the five segments tanker, dry_bulk, chemical, LNG and LPG are extracted, and all experiments are performed on them. The selection of these five segments is based on the opinion of shipping experts that the voyages of the vessels in these five segments are difficult to predict because they do not always travel between the same ports; instead, their voyages vary too much, and in the maritime industry they are the most commercially significant vessels. Therefore, they suggested that it would be intriguing to view the forecast for these vessels. After discussing the other segments' vessels, their opinion was that 'container' vessels always travel between the same ports, and it would be futile to apply any prediction model to these vessels. For 'car_carrier,' 'other,' and 'combo,' their opinion was, on these vessels, the prediction is not required because they are not the essential vessels for the maritime industry as they do no create much economic gain.

As the next step, all the voyages are been copied from the *voyages* table to the segment specific voyages table and the data is been prepared segment wise as well as the model training, validation and calculation of ETA is been done separately for all the segments.

5.1.1 Voyage definition

The voyages in the voyage table have been defined according to the definition, which was decided by the shipping experts of MO as described in section Sec-

tion 2.2.1. All the polygons for the berths and around the ports have been made manually by MO experts after analyzing historical voyages and their patterns. Their voyage definition tries to include complete voyages only, which means only including the ports on which vessel loaded and unloaded and disregarding the ports on which vessel might stop for refueling or other purposes.

Tracks builder

For making the trajectories, all the geographic points of the voyage from the departure timestamp to the current timestamp have been merged with a line, and the line has been sampled using the Douglas Peucker algorithm. Shipping experts have also verified this algorithm as this algorithm does not change the shape of the trajectory taken by the vessel for the voyage. Instead, it just reduces the excess points from the trajectory, as described in section Section 4.2.4.

5.1.2 Trajectory similarity

This feature calculates the most similar trajectory from all the given trajectories that have been departing from the same port as the matching trajectory. To compare the trajectories, a method called Symmetric Segment-Path Distance (SSPD) has been used. The SSPD method gives two values the Most Similar Trajectory's Destination (MSTD) and the distance between the two trajectories as explained in section Section 4.3.1. MSTD gives an initial prediction of the arrival port based on the trajectory similarity, and it's been analyzed that the prediction of the arrival port given by the SSPD method is also quite good. In the Table 5.1 the number of correct MSTD prediction for different segments out of total voyages has been presented.

Segment	**Total Voyages**	**Correct MSTD prediction**	**Percentage**
Chemical	210921	103451	49.04%
Tanker	810857	421321	51.96%
Dry_bulk	2005529	1020212	50.87%
lng	58486	29824	50.99%
lpg	294349	168541	57.25%

Table 5.1: Arrival port prediction result based on Symmetric Segment-Path Distance (SSPD) for different segments

In Table 5.1 the total number of voyages is the number of voyages after dividing the complete voyages into smaller voyages as described in Section 4.3.5. As

for the complete voyages, the SSPD method will give the correct results almost for all the voyages. Therefore the method has been tested after dividing the voyages. So that the method can be validated for the vessel when the voyage of the vessel is just started, in the middle, and at almost completion stage. It can be analyzed after seeing the results from Table 5.1 that SSPD was almost *50%* correct for all the segments. Therefore the benchmark has been set that ML models should predict the arrival port with an accuracy higher than the accuracies which are presented in Table 5.1 for all the segments.

5.2 Training process results

After the creation of the dataset, different ML models have been evaluated, and their hyperparameters are also tuned as described in section Section 4.4. Of all the tried models Extreme Gradient Boosting (XGBoost) model seems to have the best accuracy, so it has been selected as the multi-class classification model for this thesis. As described in the Section 4.4 the selection and development of the model is only tried on the chemical_segment. Therefore, to analyze the results on the other segments as well, the same Extreme Gradient Boosting (XGBoost) model has been trained with the same hyper-parameters separately on each of the different segments as well.

5.2.1 Data consistency

As stated in Section 4.4.2, arrival ports are eliminated to make the data consistent before it is used to train the model. The Table 5.2 shows the total number of arrival ports in the original data, the remaining number of arrival ports after the data has been consistent, and the percentage of data the remaining arrival port covers. Therefore, more than 70% of arrival ports have been removed from the dataset, and the remaining 30% of arrival ports account for more than 90% of the data in all the segments. So before training the data, the process of removing these many arrival ports has been validated by the shipping professionals. The detailed discussion related to this process has been presented in the Section 5.5.2.

Segment	**Total arrival ports**	**Remaining Arrival Ports**	**Percentage Covered**
Chemical	1503	399	92.79%
Tanker	2162	500	93.96%
Dry_bulk	3802	1025	92.87%
lng	301	120	94.3%
lpg	1109	400	95.86%

Table 5.2: Data statistics after applying data consistency steps

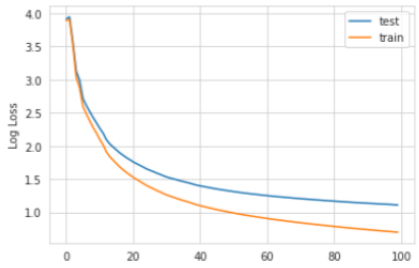
5.2.2 Loss and Error function

As the model was being trained, its performance was assessed continuously by calculating the logarithmic loss and multi-class classification error. Figure 5.1 depicts the distribution of these metrics across each boosting round in the training phase for all the segments. It can be inferred that for all the segments, by the end of 100 decision trees, graphs have been converged and almost flattened by the end. It is possible if the number of decision trees can be increased, they can converge more, but the rate of convergence is prolonged, so it might also lead to overfitting of data. It can also be seen that there is no overlapping of the train and test graph for any segment. For all the segments, test graphs stop decreasing first than the training graph. This is because the model might continue to develop to enhance its performance on the training set but not on the test set, but if the training continues, the model might become overfitted. Therefore, it is preferable to end training early and avoid overfit mode; otherwise, the accuracy on the test set will begin to drop if the model continues to train.

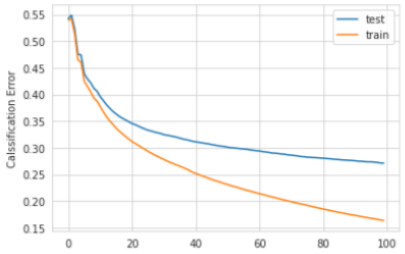
5.2.3 Feature importance

Extreme Gradient Boosting (XGBoost), a tree-based model, gives insight into the significance of features or characteristics in a dataset, an additional advantage. In a decision tree-based ensemble, the training data is evaluated to determine the appropriate attributes for creating branches while generating a tree. After training, the models can rank the features that effectively split the dataset. This is known as feature importance. In the Table 5.3 feature importance is been shown for all the segments.

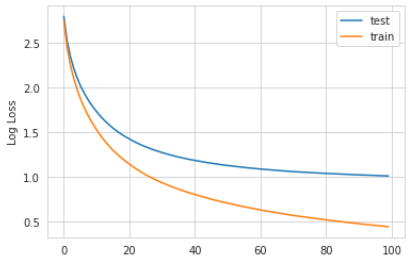
The Table 5.3 shows that sspd_mstd is the best feature for all segments, which makes sense since it is already a prediction of arrival port based on trajectory similarity. Therefore ML models are learning a lot from it. The least essential element in nearly all segments is trajectory_length, which generally provides the length of the journey, but the distance_ratio might surpass it since it almost says the same thing whether the vessel is close to the departure port or has already covered a long distance. Distance_ratio also increases the importance of the sspd_mstd feature as if the ratio is very close to '0', then the vessel will reach the sspd_mstd port as it is very near to that port. Probability is also seen to be a good feature for predicting arrival port, which could be because if the probability is high, many trajectories are going to the same port, so ML gain confidence that if the probability is high, sspd_mstd should be correct otherwise not, and sspd_mstd is already the highest contributing feature in all segments. Other characteristics such as subsegment, departure port, and sspd_dist differ across segments. The season feature appears to be not too relevant when predicting the arrival port. One reason could be that three months are grouped to define one season, and it is sporadic that a vessel takes a three-month voyage; instead, it usually goes for one or two months, therefore completing its voyage in the same season. As a result, it has little effect



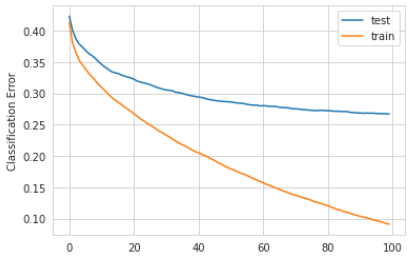
(a) Chemical log loss



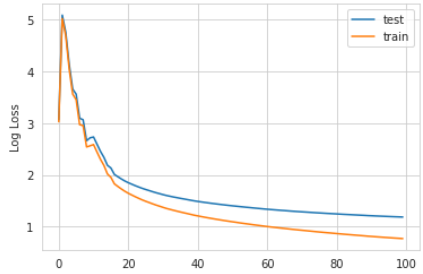
(b) Chemical classification error



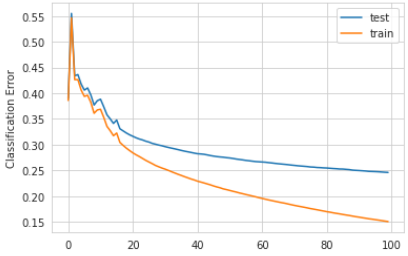
(c) LNG log loss



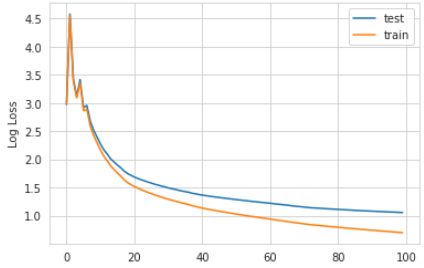
(d) LNG classification error



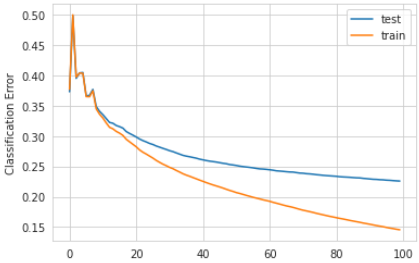
(e) Tanker log loss



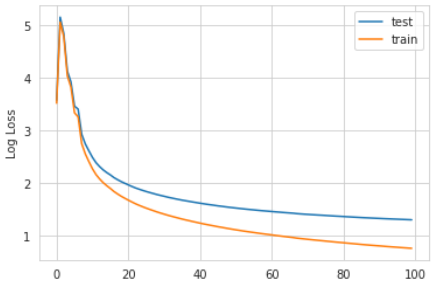
(f) Tanker classification error



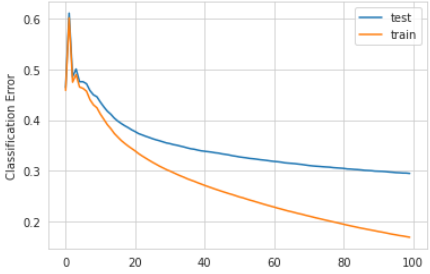
(g) LPG log loss



(h) LPG classification error



(i) Dry_Bulk log loss



(j) Dry_Bulk classification error

Figure 5.1: Logarithmic loss and classification error metrics tracked per boosting round

on the forecast.

5.2.4 Accuracy

The model's accuracy for all segments is determined using the validation dataset after training the model. The validation dataset is the 20% of data on which the model has not been trained, so it is used to calculate accuracy.

From the Figure 5.2 which shows the accuracy for all the segments, it can be analyzed that the model performs best for the LPG segment and the worst for the dry_bulk segment.

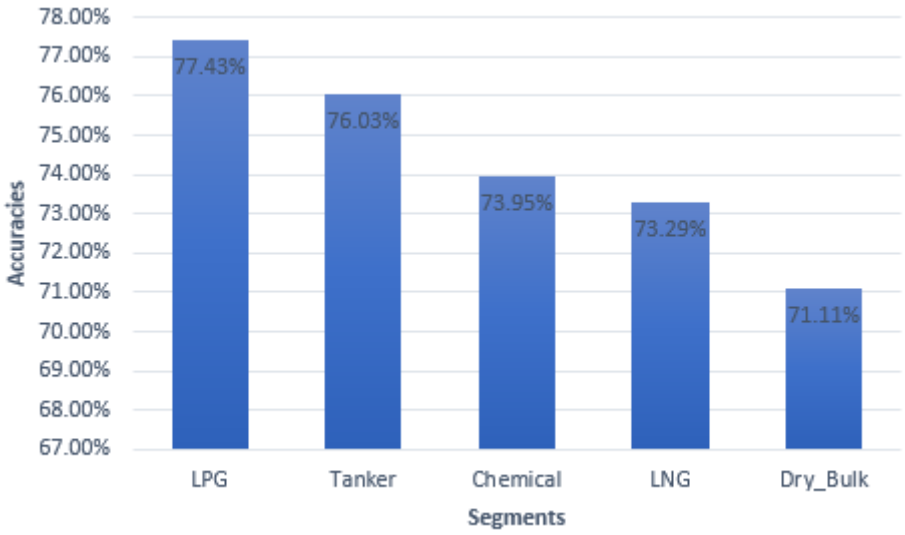


Figure 5.2: Accuracy of prediction for different segments

5.3 Results for availability of vessel at a port

The steps for predicting vessel availability at a port have been done in two parts, as discussed before. The first one is to predict the arrival port for all the vessels which are in the middle of their voyage. The second part is to calculate the Estimated Time of Arrival (ETA) to the expected port for all those vessels.

5.3.1 Prediction of arrival port

For the first phase, to select all vessels in the middle of their voyage, all vessels that started their voyage before February 1, 2022 but completed after, are chosen for each segment. The previously trained ML model for all segments only saw voyages that ended on or before February 1, 2022. However, the test data include all

Feature	Importance
sspd_mstd	0.646544
probability	0.158699
departure_port	0.076167
sspd_dist	0.049959
sub_segment	0.024121
distance_ratio	0.023284
season	0.016584
trajectory_length	0.004641

(a) Chemical Segment

Feature	Importance
sspd_mstd	0.369643
sspd_dist	0.286682
departure_port	0.135356
sub_segment	0.067184
probability	0.050609
distance_ratio	0.049544
trajectory_length	0.036392
season	0.004591

(b) LPG Segment

Feature	Importance
sspd_mstd	0.427103
sub_segment	0.256165
probability	0.099650
departure_port	0.099358
sspd_dist	0.032856
season	0.031913
distance_ratio	0.030549
trajectory_length	0.022407

(c) LNG Segment

Feature	Importance
sspd_mstd	0.404957
sspd_dist	0.281838
probability	0.169182
distance_ratio	0.079442
sub_segment	0.029749
departure_port	0.025565
season	0.006054
trajectory_length	0.003213

(d) Tanker Segment

Feature	Importance
sspd_mstd	0.419284
probability	0.218736
sub_segment	0.171288
departure_port	0.074156
sspd_dist	0.058143
season	0.041896
distance_ratio	0.013264
trajectory_length	0.002963

(e) Dry_bulk Segment

Table 5.3: Feature Importance for all segments

journeys completed after February 1. So the model's accuracy will be assessed on previously unknown data, therefore, demonstrating how the model will perform on real-world data when implemented in production.

To predict the arrival ports from the model, the data should have all the features as the training data. Therefore all the features have been calculated and added to the test data. But for the test data, no vessel voyage has been divided into shorter voyages, it was done only on the training data so that models can be trained on all the lengths of voyages. The vessels are already in the middle of their voyages for the test data.

On test data, the sspd_mstd ports have been assessed to check the number of correct predictions made by the trajectory similarity method. In the Table 5.4 the total number of vessels selected that fit the criteria of test data are shown, the correct predictions made by the SSPD model on these vessels' voyages and the percentage of the vessels predicted correctly by SSPD method.

Segment	**Total Vessels**	**Correct MSTD prediction**	**Percentage**
Chemical	311	73	23.47%
Tanker	1320	409	30.98%
Dry_bulk	5889	2533	43.01%
lng	172	60	34.88%
lpg	486	154	31.68%

Table 5.4: Arrival ports prediction based on Symmetric Segment-Path Distance (SSPD) method on test data

The next step is to make the test data ready to be predicted by the ML model. Similar to what was discussed before in the Section 4.5.2, the ML model will need the exact mapping of categorical columns exactly as the training dataset. As a result, it is essential to ensure that all categorical features in both the training and test datasets must have the same classes. Among all the categorical columns, 'season' and 'sub_segment' have all their classes maintained, but for 'departure_port' and 'sspd_mstd,' the classes must be removed from the test data. Table 5.5 describes for all the segments the total number of data points that have to be removed based on departure_port and sspd_mstd that cannot be predicted by the ML model.

From the Table 5.5, it can be seen that only a small number of vessels have been eliminated, which further confirms that it was acceptable to remove the arrival ports from the training data, which accounts for a significantly less percentage of data, for training purpose. To make the data more consistent, so that the

Segment	Total vessels	Removed departure_port	Removed sspd_mstd	Remaining vessels
Chemical	311	17	7	287
Tanker	1320	20	10	1290
Dry_bulk	5889	33	23	5843
lng	172	2	2	168
lpg	486	6	4	476

Table 5.5: Number of vessels removed based on departure port and sspd_mstd that cannot be predicted by ML model

ML model performs well.

The last step is to apply the ML model to the final data and forecast the ports of arrival along with the probability of reaching each predicted port. As discussed before, for those vessels whose arrival port is not predicted by the ML model, they are assigned the same predicted port as the sspd_mstd because that is also a good prediction. For the probability, they are assigned the same probability as the feature set probability because it also indicates the possibility of the most similar trajectory among all similar trajectories reaching the MSTD. After integrating the results from the sspd_mstd and ML models, the Table 5.6 displays the number of accurately predicted arrival ports for all segments.

Segment	Total Vessels	Accurate Predicted Vessels	Accuracy
LPG	486	301	61.93%
Tanker	1320	792	60.00%
Chemical	311	193	62.05%
LNG	172	108	62.79%
Dry_Bulk	5889	3477	59.04%

Table 5.6: Accuracy of different segments on the test data

It may be deduced from the Table 5.6 that the accuracy of all segments is considerably decreased. The Table 5.4 shows that the prediction made by the SSPD is also decreased, and ML model had the greatest feature importance of sspd_mstd to forecast the right arrival port, as seen in Table 5.3. The sspd_mstd accuracy might be decreased because there might be a possibility of having a small length of trajectory covered by the vessels till February 1st, 2022, and SSPD methods seem not to perform well for the small length of trajectories. Therefore, the accuracy of the ML models on the test data seems to be lower than the training accuracy.

Figure 5.3 shows the accuracy for all the segments on the test data.

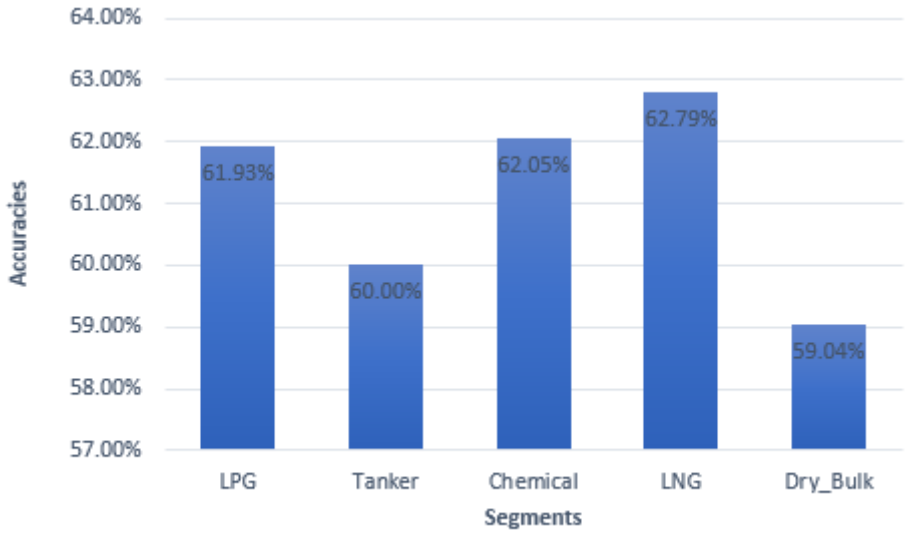


Figure 5.3: Accuracy of prediction for different segments on test data

5.3.2 Data analysis on predicted data

Data analysis has been done on the final predicted data to get more insights into the predicted data. Some of the results of the data analysis are discussed below:

- The mean error distance between the erroneous projected ports and the actual arriving port has been determined. Figure 5.4 shows the actual port and the predicted port both are in Japan, but they have been predicted wrong. So the Haversine distance is calculated between these two ports. This information will be utilized to know how distant is the expected port from the real port. If the distance is small, then the projected port is exceptionally close to the actual port, but the anticipated port is nowhere near the actual port if the distance is enormous. According to the shipping experts, if the area is also appropriately predicted, then also it is useful. If the error distance is close, the projected port is erroneous, but it is very near the actual port, so it is still a good prediction. In the Table 5.7 for all the segments mean error distance has been listed.

  Figure 5.5 shows the graph for the error distance varies according to the number of vessels which are predicted wrong for the chemical segment. From the Figure 5.5, it can be seen that only a small number of vessels account for maximum error distance, while the maximum number of vessels have very low error distance. So it can be inferred that although the vessels

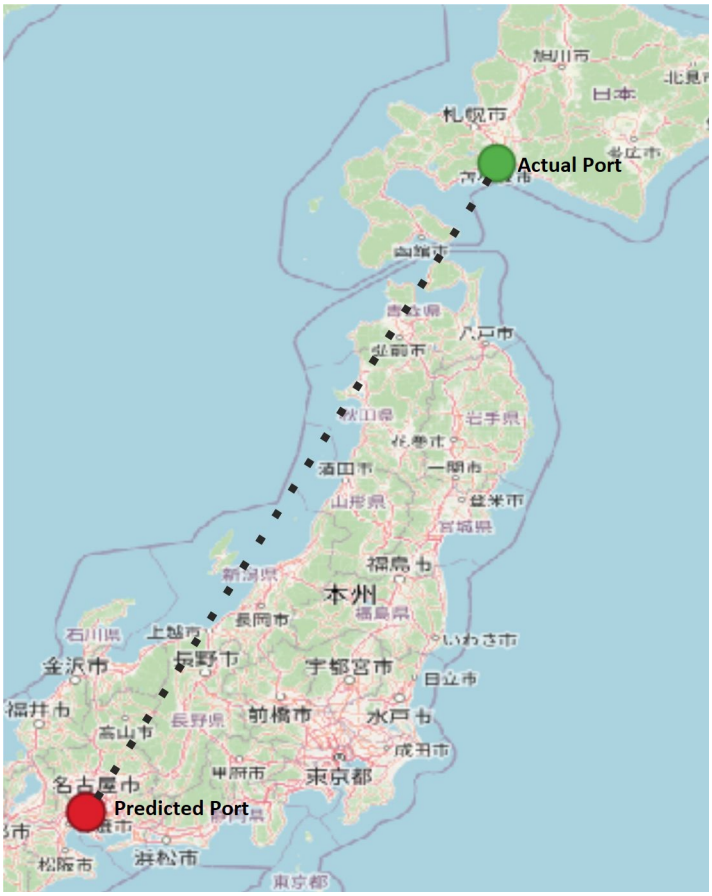


Figure 5.4: Distance between incorrect predicted Port and actual Port

Segment	Mean Error Distance (in Km)
LPG	2098.23
Tanker	2871.11
Chemical	2021.89
LNG	1992.99
Dry_Bulk	2097.67

Table 5.7: Mean error distance between incorrect predicted port and actual port

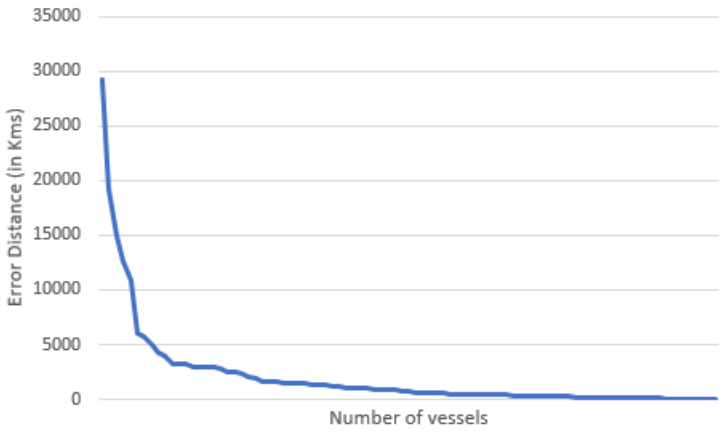


Figure 5.5: The variation of error distance according to number of vessels

are predicted wrong still, they have been expected in the right geographical area by the models.

- The average probability for the accurately predicted port and the incorrectly predicted port has been determined. The mean probability is more significant for successfully predicted ports, it also seems reasonable as the ports which are accurately predicted by ML models or SSPD models should likewise have high probabilities as the models are sure for the prediction. And for the incorrectly predicted ports, the mean probability is lower. This indicates that the ML model predicted the port, but it was uncertain. Thus, the probability is lower, and the port is also inaccurate. The Table 5.8 gives the average probability for accurate and wrong predictions for each vessel segment.

Segment	**Average Probability for correctly predicted ports**	**Average Probability for incorrect predicted ports**
LPG	61.95%	35.96%
Tanker	55.98%	40.77%
Chemical	60.91%	31.96%
LNG	71.87%	30.48%
Dry_Bulk	57.83%	39.27%

Table 5.8: Probability across different segments for correctly predicted ports and incorrectly predicted ports

- Combination of both models have been accurately predicting ports, but it will be fascinating to see how well the model predicts countries. As the first two characters of UN/LOCODE for ports correspond to the country, which is also specified in Section 2.3.1, therefore, it is possible to check for countries additionally. The data in the Table 5.9 demonstrates that countries are more accurately predicted than the ports. It is also logical since the countries have higher granularity than ports; there might be several ports inside a country, but all the ports are now aggregated into a single country. Therefore the accuracy should be greater. The Table 5.9 displays the accurate country predictions for each vessel segments.

Segment	**Total Vessels**	**Accurate Predicted Country**	**Accuracy**
LPG	486	382	78.60%
Tanker	1320	967	73.25%
Chemical	311	269	86.49%
LNG	172	139	80.81%
Dry_Bulk	5889	4829	82.00%

Table 5.9: Accuracy for the correctly predicted countries

5.3.3 Estimated Time of Arrival(ETA)

As discussed before, the calculation of Estimated Time of Arrival (ETA) has been done using the routing engine which was given by Maritime Optima AS (MO) which calculates the ETA. And after the calculation of ETA the final table is been made with the details of IMO, departure_port, predicted_port, predicted_probability, and the ETA.

For example, the vessel owner wants to know how many vessels will be there at port 'JPCHB,' the most popular port in Japan for chemical cargo, to forecast the supply and demand. If the supply matches up with the predicted demand, there is no point in sending the vessel to that port. On the other hand, if there is more demand than cannot be met by the expected supply, there will be value in sending the vessel. So from Table 5.10 it can be analyzed that there are many vessels that will arrive today, and there is only one vessel that will arrive after two days. Therefore the chemical vessel can be sent after one day as there will be no vessels at the port after 24 hours and before 48 hours.

imo	departure_port	predicted_port	predicted_probability	ETA (in hours)
9809289	JPSMZ	JPCHB	87.18%	4.219
9677064	JPMIZ	JPCHB	79.49%	9.065
9677193	JPANE	JPCHB	30.41%	1.725
9606986	INVTZ	JPCHB	77.16%	1.744
9466001	USHPY	JPCHB	99.72%	0.453
9912476	JPOSA	JPCHB	17.72%	9.566
9675341	JPKSM	JPCHB	40.66%	47.921

Table 5.10: Chemical vessels which can arrive at port of 'JPCHB'

5.4 Adaptations after analysis of results

After observing and analyzing the results, various additional things were done to see if they might enhance the outcomes. In this part, the things that have been attempted are mentioned. All of these adaptations were carried out on a single segment, which is chemical; if the accuracy is increased and the modification seems to be important, the adjustments may be applied to the other segments as well.

5.4.1 Countries prediction

As can be observed, to keep the data consistent before training the ML models, most of the information is eliminated, causing the ML model to be trained on just a few arrival ports, resulting in only a few of arrival ports can be predicted in the future by the ML model. One adjustment has been implemented to address this problem: rather than anticipating ports, it is simpler to forecast countries. As explained in section Section 2.3.1, the first two letters of UN/LOCODE reflect the country name. Therefore, identifying the nation from the ports is feasible.

As previously stated, there are 1503 distinct arrival ports for the chemical segments. However, when the countries are considered, there are only 139 distinct countries. So the number of classes to predict is dramatically decreased, and the data was still inconsistent, but rather than deleting any data points, the model is trained on the whole dataset. In addition, the departure ports and sspd_mstd have been relocated to countries from the ports. As a result, the journey is defined as the vessel leaving one country and arriving in another.

For training, the same ML model, Extreme Gradient Boosting (XGBoost), is used; however, the model is trained using the XGBoost default hyperparameters. The validation dataset, which included 20% of the total dataset on which the model had not been trained, achieved an accuracy of 89%. This seems to be pretty

high compared to the port accuracy and that too on the whole dataset without removing any data.

Shipping experts have evaluated this change, and they conclude that it enhances accuracy, which is a positive thing, but they have expressed one issue about it. The difficulty was that anticipating the country was worthless for more major countries with hundreds of ports, such as the United States and Russia; instead, knowing the precise port was required. They also mentioned that developing the model with the smallest granularity is desirable. Even if the model does not predict well with the smallest granularity, this is acceptable since when the port is forecast, the nations can also be predicted from the ports, but not the ports can be known from the predicted countries. As a result, this adaptation was abandoned, and no more research was conducted.

5.4.2 Training on full dataset

Many ports are eliminated throughout the training phase to ensure consistent data. Therefore the model was tested with the whole dataset without removing any data points. Over the entire dataset for the chemical section, the same ML model with the same hyper-parameters was attempted. The accuracy was reported to be 59%. It is plausible that it can be improved when the model is tested with alternative hyper-parameters tuned based on the whole dataset. Still, it is improbable that the model will provide the same accuracy as the consistent data accuracy. Furthermore, as advised by the experts in the interview part Section 5.5.2, it is unnecessary to train the model on the whole dataset; the most frequent ports are sufficient to forecast. As a result, this modification has been rejected, and it has not been tried on the other segments.

5.4.3 Group sub_segments

The sub-segments are classified according to the size of the vessel and the weight of the cargo they transport. As a result, there are eight sub-segments for the chemical segment based on the size of the vessel and the weight of the cargo vessel carries. It has been seen that instead of eight classes, the vessels can be grouped into three classes which can be defined by: small, medium, and large. So, after grouping the sub-segments, the model is trained, and accuracy is calculated. The feature significance is also determined to check whether the sub-segment feature plays a more significant role now in predicting the arrival port after grouping.

The grouping of sub-segment for the chemical vessels is been defined in Table 5.11. The process of which sub segment belongs to which group is been decided by the data analyst of MO.

There was no change in accuracy for the chemical segments when the Extreme Gradient Boosting (XGBoost) model was used after grouping eight sub-segments

Group Name	Included sub segments
small	small_1, small_2, unspecified
medium	handy, flexy, intermediate
large	medium_range, panamax

Table 5.11: Group of sub segments for the chemical vessels

into three. It was precisely the same as 73.85%, in fact, reduced by 0.1%, and the feature significance remained unchanged from what was mentioned in the Table 5.3 for the chemical segment.

Following a discussion of the aforementioned adjustment, experts said it is an excellent effort. Many of the vessels are predicted by their size, as small vessels only go for a short distance, but huge vessels travel for longer distances. The Figure 5.6 displays the journey projection depending on the size of the dry_bulk segment. The yellow arrows indicate smaller vessels, and it can be seen they are restricted to areas such as Europe and China, but the red arrows symbolize larger vessels and have longer trips that begin in China and go through Africa to Brazil. According to experts, segment grouping is not very relevant for the chemical segment because they are very unpredictable, but for the dry_bulk and tanker segments, a group of sub-segments will improve the accuracy, and the feature importance will also be higher for the sub-segment while predicting arrival ports by ML models.

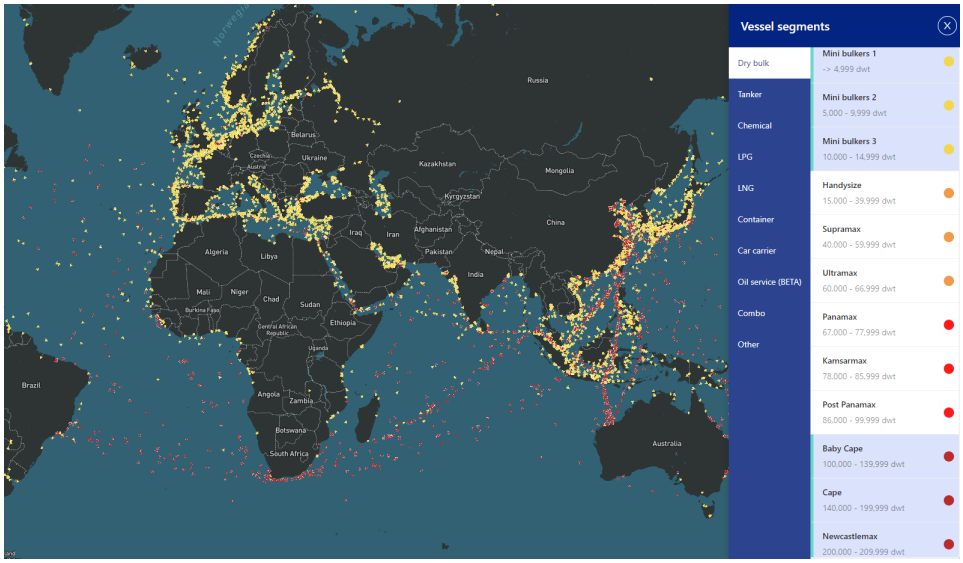


Figure 5.6: Variation in voyages based on the size of vessels

5.5 Experts interview

To get validation of the suggested solution and to determine the practical use of the solution offered in the thesis. The participating firm Maritime Optima AS (MO) contacted the shipping specialists. Three shipping specialists were interviewed, each specializing on a different aspect of the shipping industry. One of them is a shipping data analyst, another is a broker in the shipping sector, and the final one owns a vessel and works as a vessel owner. The suggested solution, including all steps performed, is presented to them, and they are then asked to give their view points on the following topics:

- What are the existing solutions available to predict the vessel availability, and what are the limitations of those existing solutions?
- Why do only some ports consist of so much data while the majority have only a few data points, so it will be OK to train the model only for a few arrival ports?
- If this solution is commercialized, what will the commercial value or practical gain it can provide?
- What different things or modifications in the presented model will be recommended to increase the study's commercial gain in the future?

5.5.1 What are the existing solutions available to predict the vessel availability, and what are the limitations of those existing solutions?

The answer to this question was the same from all the interviewers. The answer was almost the same as discussed in the Section 1.3. At present, all the information on the availability of vessels is gathered through contacts. The ship owners and the cargo owner are dependent on the brokers for the information. And the brokers get the data from the other brokers and other various ship owners to make a detailed report of all the vessels' present position and their availability at the ports. The exact answer of one of the interviewees is:

Shipbrokers provide this information to ship owners and cargo owners by phone or email. The shipbrokers' contribution is that they phone the ship owners and inquire about the status of their vessels. They then compile that information and generate position lists, which they email all vessel owners and cargo owners (their customers). Many shipbrokers provide the data they gather in Excel spreadsheets so that it can be shared with others. Then, some brokers used their software, and some of the largest shipbrokers began to investigate AIS. Then certain software businesses, such as Kpler[1] and MO, have started to give real-time vessel information through the high class visualization sofwares which are based on the AIS data.

[1] https://www.kpler.com/

The biggest shortcoming of this strategy, according to the interviewers, is that the process is not arranged systematically. According to interviewees, no broker can save information for all vessels for a single segment. Making a complete report for all vessels after obtaining a lot of information via phone calls and emails is a huge task. Furthermore, the report cannot be fully trusted since, according to the interviewer, ship owners and other brokers sometimes supply incorrect information or sometimes do not provide it at all. As a result, present models rely on physical interactions, and there is no infrastructure available to give data on vessel availability. As a result, there is no transparency in the current system; all actors rely on one another for knowledge, and information between them is transferred for a fee or not shared.

5.5.2 Why do only some ports consist of so much data while the majority have only a few data points, so it will be OK to train the model only for a few arrival ports?

For this question, all interviewees are provided with the findings that are presented in the Table 5.2 and asked about the reasons and validity of having just specific arrival ports for the majority of the data and majority ports for so little data. The initial response from all interviewees is that it is reasonable to eliminate the arrival ports that appear just a few times in the database. According to one of the interviewers, one of the reasons may be that the data is from December 2019, and some vessels may have been chartered on Time Charter (hiring of vessels for a specified period) to some operators who might be traded the vessels in specific trade only during that time. So there are numerous voyages between certain ports and relatively few between others. According to another interviewer, just a few of the world's ports handle more than 90% of maritime tradings. The MO has already filtered out 5342 of all the world ports, but according to the interviewer, only half of those 5342 will handle significant trade. So, if the thesis model can also forecast those ports, it is equally significant since the other ports are not that important.

Furthermore, the interviewers think that the final prediction is also based on the trajectory similarity method. Therefore it is essential to train the ML model only on the most frequent ports as the prediction accuracy should be higher. It does not matter if it predicts less number of ports, and the ports that the ML model cannot predict will be predicted by the trajectory similarity method, which according to interviewees, is also a good prediction. So in the final, there will be predictions for all the vessels with all the ports, which according to all of them, is an excellent approach to predict the high occurrence ports by ML model and the others by trajectory similarity method. Therefore, there will be a prediction for all the vessels in the end.

5.5.3 If this solution is commercialized, what will the commercial value or practical gain it can provide?

The answer to the questions is asked from the point of view in their respective fields. As there were interviewers belongs to three different sectors of the shipping industry so the answer are basically based that how the solution will help for their specific sectors. Their answers are as follows:

Shipping data analyst

According to the shipping data analyst, the solution will aid in the systematic understanding of the supply situation in real-time. According to the interviewer, if there is a strong demand for vessels but no supply, the market value would be high. The projection of vessels that may be able to satisfy demand might potentially disrupt the market and result in lowering the overall market value. It is also explained with an example: the interviewer said that there is a large amount of oil to be carried from the Middle East, but there is no expectation on vessel availability. So the market will be pretty high since demand is extreme, but supply is uncertain. However, if the model predicts and also indicates that several vessels are due to arrive in the Middle East in a short period, the market value would plummet. The ship's identification might potentially cause market disruption. And according to the interviewer, it is not necessary to anticipate the precise port or country; if the area, for example, the Middle East, can be predicted for vessel availability, it is sufficient to alter the market. As a result of the data analyst's conversation, the model may be trained to anticipate the availability of vessels in the area. As it seems that the accuracy was good in the case of country prediction, the model will likely perform better in the case of regions. Therefore according to the data analyst's point of view, this solution will help to understand the market situation in real-time systematically.

Shipbroker

According to the shipbroker, the brokers need reliable information to provide to their customers, which includes cargo owners and vessel owners. Therefore, the forecast may be secondary to the information received by communication with different vessel owners and other brokers. According to the broker, the estimates for the availability of the vessels would undoubtedly assist, as it will organize everything so that all the data for all the vessels' prospective availability at a port can be viewed. Furthermore, it is not feasible for a single broker to get all vessel information for a single segment. So, with the solution, there will be statistics about all the vessels and where they will be at a given moment. However, with this solution, exact information is also required. So, according to the interviewer, the solution of information gathering through contacts and the model predictions can be combined to know about all the vessel's availability for a cargo, therefore a

more robust report can be prepared to supply to the clients, such as vessel owners and cargo owners.

Vessel owner

According to the vessel owner, this approach will significantly aid in scheduling the fleet of vessels efficiently. According to the vessel owner, after the ship has emptied its cargo at the unloading port, it must proceed to the loading port to take up cargo. So, for vessel routing, ship owners rely on brokers, who assist them in meeting with cargo owners and finalizing an agreement to convey their cargo next, so the vessel is routed accordingly to the cargo loading port. Furthermore, there is a lot of negotiating in the price, as the cargo owner wants to transport the cargo at the lowest possible cost, so there is bidding between all the vessels that can be available at the port for taking the cargo and whoever bids the least, that vessel owner's vessel will be responsible for transporting the cargo. So, according to the interviewer, the solution will bring transparency to the system; now, information is only available by brokers about the number of vessels competing for cargo, but with the solution, data on the number of vessels that may compete for cargo at a specific port can be known. Due to system transparency, the vessel owner will now have access to information on all of the vessels availability at port for that cargo. Furthermore, the solution will help the vessel owners to route the ship to the port where fewer vessels are vying for cargo. As a result, there will be efficient planning of the vessels, resulting in high revenue from the vessels.

The shipowner also said that the owners are more concerned about the routing of larger vessels. Therefore, if the solution's accuracy can be increased simply by concentrating on the larger vessels within a segment, it will bring a more significant commercial advantage. Therefore in the future, a model can be tried with the large vessels only, and the accuracy can be tested. Finally, the vessel owner was very excited and wished to see the working of the solution in the production as soon as possible as it could be instrumental in increasing the revenue.

5.5.4 What different things or modifications in the presented model will be recommended to increase the study's commercial gain in the future?

The last question, interviewees were asked to recommend any improvements that may be utilized to enhance the solution further. According to one interviewee, if the model for predicting arrival port can be trained for large vessels only within a segment and it can provide more accuracy, than the solution will be more valuable. Because smaller vessels mostly take short voyages, their information is irrelevant to shipbrokers or owners. However, big vessel information is critical since their travels may be lengthy; therefore, shipping professionals are more interested in large vessel information than small ones as they carry more value.

The other interviewer believes that knowing whether the vessel is loaded or empty may increase accuracy. Because loaded vessels will more likely go to unloading ports and vice versa, the issue with this feature is that there should be data for port classification, such as whether the port is a loading port, an unloading port, or a bunkering port. A port may also perform several functions, such as loading and unloading. However, the interviewer believes that if this information is accessible, it will significantly influence the forecast.

One of them also mentioned that the routing engine developed by MO and utilized in the thesis did not account for weather conditions, and the vessel's path is also affected by weather circumstances. In severe weather, the vessel is compelled to deviate from the scheduled course and choose an alternate route, causing the ETA at a port to vary. Therefore in the future, the calculation of ETA will give a more accurate prediction of the availability of the vessel if calculation also considers weather conditions.

5.6 Results conclusion

This section shows that the XGBoost model performs best on the LPG training data with 77% accuracy, but when tested on actual unseen data, it performs best on the LNG training data. It is also noted that when the granularity is raised, the model seems to function well, as the model accuracy on the chemical segment goes from 73% to 89%. Furthermore, it is determined that based on the ports that were eliminated during the training the model, there was not much data in the test data for that certain ports, which is a good thing and also supports that vessels generally go less between those ports. Furthermore, the probabilities of successfully predicted ports are considerably greater than those of mistakenly predicted ports. Finally, the interview with specialists from various parts of the shipping business reveals that there was no transparency in earlier ways of knowing port availability. The information being conveyed was based on phone conversations and emails between vessel owners and brokers. As a result, this strategy provides transparency while also assisting in systematically analyzing supply-demand at various ports. Some experts believe that the model should concentrate more on larger vessels within a segment and enhance accuracy to be more commercially beneficial. The next chapter will summarize the whole thesis, including potential solutions to the research questions and thesis constraints, and future studies.

Chapter 6

Discussion

In this chapter, a summary of the thesis will be included, along with a discussion of the research questions and, in the end, some of the thesis's limitations.

6.1 Summary

The thesis is done with the collaboration of maritime company Maritime Optima AS (MO). They give the project title and problem, and according to them, the solution of predicting the vessels' arrival at port for a specific cargo can be very beneficial commercially for the maritime industry workers. Therefore an attempt was made in the thesis to find the solution to the problem. In the end, the solution's validity and the commercial applicability are determined to exist after the solution is available on the market.

For the thesis, two primary objectives needed to be fulfilled. The first one was to study the existing solution for the prediction of arrival port and then find a better solution that could give better results. Another one is to see if there is any research on the availability of vessels at a port that is too specific for a cargo. If the answer is 'yes,' then try to provide a better solution; if 'no,' then find a solution and validate it from the shipping experts.

6.1.1 Prediction of arrival port

There were many studies which were focusing on the prediction of arrival port. But the major drawback with all the studies is that they mainly focus on a certain geographical area, not covering the whole area. Only a few studies [5, 17] found out that focused on the entire geographical area. So these studies have been studied to find out the methods and features they have been using to predict arrival port. The primary research was the thesis [5] which was done in the same company as MO, so some features have been taken from that study are Most Similar Trajectory's Destination (MSTD) which was calculated using SSPD method, and

vessel information which are segment and sub-segment to which belongs.

Following the selection of features, the creation of the dataset begins. The MO provides the 'ports,' 'voyages,' and 'AIS' tables that are required for the development of the dataset. The development of the dataset began with the production of tracks for all voyages, which was accomplished by combining all the location points with a line to produce a trajectory, which was then sampled using the Douglas Peucker method. All other features were then calculated, which are: sspd_mstd using SSPD technique, the season based on the departure timestamp, probability, and distance ratio. The dataset also includes the sub-segment to which each vessel belongs, which is already supplied by the MO in the voyages table. Additionally, the voyages for the training set have been divided into smaller voyages so that the model can be trained on voyages of varying lengths.

The construction of the dataset is followed by the training of the model for the prediction of the port of arrival. Before training the model, the data is visualized and found to be quite inconsistent; thus, arriving ports that account for a smaller fraction of the data are eliminated from the dataset to make the data consistent. The removal of the arrival port occurs in two steps; in the first step, the combination of arrival port and departure port is examined, and the arrival port that does not account for nearly 90 percent of the combination is removed; in the second step, only arrival ports are examined, and the arrival port that accounts for almost 95 percent of the data is kept; other arrival ports are removed. This step of data consistency has been validated by the shipping experts also.

The Extreme Gradient Boosting (XGBoost) model has been selected for model training, followed by tuning the model's hyper-parameters and selecting the optimal hyper-parameters for training purposes. The same model with the same hyper-parameters is trained independently on the five segments chosen for this thesis, namely chemical, LNG, LPG, tanker, and dry_bulk. They have received specific training since the challenge in the thesis was the availability of vessels for a particular cargo. Consequently, if the cargo is chemical, the chemical vessel prediction can be made using the model trained on the chemical dataset alone; similarly, the models can be applied to the vessels necessary to carry the specific cargo.

From the results for the prediction of arrival port following things can be analyzed:

- The sspd_mstd prediction was accurate for almost 50% of voyages.
- For all segment only 30% of arrival ports accounts for more than 92% of data.
- After analysis of loss and classification error graph from Figure 5.1 it can be ensured that there was no overfitting of the model.
- The model accuracy performs best for the LPG segment, and worst for dry_bulk.

- Analysing the feature importance from Table 5.3 sspd_mstd accounts for the maximum importance for all segments and the trajectory_length for the least.

6.1.2 Prediction for availability of vessel at port

No prior research was discovered that focuses solely on predicting the availability of vessels at a port for a certain cargo. There was research involving the forecast of arrival port or the prediction of ETA, but no study addressed the problem of vessel availability. Therefore, a technique is developed in this thesis that can solve the issue by integrating two solutions, namely the prediction of the port of arrival and the calculation of the ETA to the anticipated port. By combining these two techniques, the availability of a vessel for a specific cargo can be determined.

The initial need for predicting vessel availability is to identify all vessels in the middle of their voyages. To satisfy this condition, a date has been chosen such that all vessels may begin their voyages at any time before this day but must finish their voyages after this date. Accordingly, the date chosen is February 1, 2022, since the training of the models for each segment has been completed for voyages accomplished by that date, and the models have not been exposed to data for voyages made after that date. Consequently, using this data, the models will also be evaluated on wholly unobserved data, and the availability of vessels at a port can be predicted.

Following the selection of vessels for each segment, the ML models are used to estimate the ports of arrival for each of these vessels. To predict the arrival ports by the ML models, there are certain vessels for which the ML models cannot predict the port due to the removal of many data points during the training process of the model. Consequently, these data points are deleted from the dataset, and predictions are made for the remaining vessels, along with the probability for the predicted port. For vessels eliminated from the dataset that the ML model does not predict, the predicted port and probability are assigned the same as for the sspd_mstd and probability value, respectively, which was determined during the creation of the features. Therefore, the final projected table comprises predictions from the ML models and the SSPD based trajectory similarity approach.

After the prediction of arrival ports and the probability for all the segments by the combination of their specific ML models and SSPD based trajectory similarity method, the last step is to calculate the ETA to the predicted port. For the computation of ETA, the routing engine of MO is used, since it is the best-in-class routing engine and provides the ideal time to reach the specified location. The current location coordinates, the projected port coordinates, and the vessel's speed are supplied to the routing engine, which calculates and returns the ETA to the predicted port. The value of ETA, along with the departure port, the IMO of the

vessel, the expected port, and the forecast probability, are stored segment-wise in the final tables so that it is possible to predict which vessels will arrive at the port for a specific cargo.

From the results for the prediction of availability of vessels at a port for the specific cargo following things can be analyzed:

- The accuracy of ports predicted by the sspd_mstd was not as high as for the training data, this can be due to the reason that there might be only short voyages cover till February 1st, 2022 by the vessels.
- The number of vessels removed that cannot be predicted by the ML was significantly less for all the segments. This signifies that only a few vessels travel between the ports that were removed during the training process. Therefore, removing some data points to make the data consistent during training is justified.
- Accuracy on the test data after the combination of ML model and SSPD method is not as high as on training data. But this can be justified by the low accuracy percentage of SSPD method.
- The probability of correctly predicted ports was in the range of 60%. So if the ports are predicted correctly, the model was assured of the prediction made.
- The countries' accuracy is high for all the segments, and it is expected as the countries have higher granularity than ports so that they will be predicted more accurately than ports.

6.1.3 Additional steps performed

As a consequence of the study of data, several alterations are implemented to improve outcomes. The initial change was to anticipate countries rather than ports since countries are more consistently distributed on data than ports. The accuracy was also more remarkable for the prediction of countries. Still, shipping professionals did not accept it since the country's forecast is not advantageous for larger countries. The second adjustment was to test the model on the whole dataset to evaluate the accuracy with and without data removal. It was seen that without reduction, the accuracy was relatively poor, and shipping specialists also recommended removing the data points. The third adjustment consisted of grouping sub_segments to see whether the accuracy changes and the sub_segment's feature relevance will be given more weight in the prediction by the ML model after the alteration. But no change was seen in the chemical segment. But according to experts, it can be tried on dry_bulk as it has many different sub-segments and the voyage of dry_bulk also vary according to sub_segments so that sub_segment groups will play an essential role in the case of dry_bulk and tanker.

Finally, the whole thesis was presented to shipping specialists, who validated the various procedures described in the thesis. In addition, the experts were asked

to specify the value that the thesis solution would provide to the maritime industry.

6.2 Research questions

This section describes how each research question was addressed by the recommended solutions presented in the thesis which were mentioned in Section 1.5.

6.2.1 RQ 1: How can the AIS data be used to predict the future destination of the vessel?

Analysis of previous studies shows that, for all of the research work done in the marine business, there was just one data source: AIS data. As indicated in the Section 2.1.1 AIS data, provides location , speed, and other vessel attributes such as size, type, and IMO number. As a result, additional significant information may be retrieved from this data, such as the route taken by the vessel to reach the destination port from the departure port by linking the vessel location data supplied by the AIS. In addition, the kind of vessel may be established from the size and type of cargo carried from the data present in the static field of the AIS data. Furthermore, the speed of the vessel and the course and heading are all contained in the data. Previous research has shown that these characteristics may be used to predict the arrival port.

However, since most AIS data is erroneous, a significant amount of processing is necessary before it can be utilized to forecast the arrival port or predict the ETA. The AIS data used in this thesis was given by the partnering firm MO; hence the data used in this thesis is the data that was utilized after all of the pre-processing processes that the shipping specialists took to clean the data. As a result, the data in the thesis can be trusted. Furthermore, if any irrelevant data is present, it has been eliminated, such as some of the voyages, as indicated in the Section 4.2.3.

In this thesis, the voyage definition has been defined utilizing the AIS data since when defining the voyage, navigational status and speed data are required which are provided by the AIS data. The tracks were generated by combining AIS positional data. Other features such as Most Similar Trajectory (MST), season, probability, and distance_ratio are constructed solely from data tables generated by extracting and processing essential information from AIS data. Therefore AIS data is the source of all the developed features which are used in the thesis for the creation of a solution.

RQ 1.a: What kind models and data have been used to predict the destination of the vessel?

Several models and data characteristics have been utilised in the prediction of arrival port. However, most research papers have formulated the problem of fore-

casting the arrival port as the classification problem and solved using conventional classification tree-based methods such as Random Forest (RF) models or Extreme Gradient Boosting (XGBoost) models. In addition, some articles employ the sequence to sequence modeling, but the accuracy of the arrival ports was relatively poor or not indicated, as mentioned in Section 3.1. Furthermore, the authors of the research [21] demonstrate that in solving the issue of arrival port prediction, the classification models, especially the tree-based, yield the best results.

If data is analyzed that has been utilized for prediction, it is AIS data for all research. For example, the articles [5] demonstrate the significance of vessel type, whereas the papers [5, 17] illustrate the importance of incorporating prior trajectories when forecasting the arrival port. Course, heading, and draught are additional vital factors that the writers of [18, 19, 21] employ when estimating the arrival ports. However, as indicated in the Section 4.3.6, the data of cpurse, heading and draught is mainly inaccurate. For these papers, they have been tackling the issue that has been supplied in a challenge; thus, the data is restricted, and these figures may be accurate for that limited data. However, when compared to actual data, these numbers are typically incorrect and are not recommended to be used for prediction.

However, most of the investigations that have been conducted are geographically limited. Therefore, the only global research discovered is [5, 17]. Both of these works consider prediction a classification issue and use tree-based techniques for prediction, with [17] using the RF model and [5] employing the XGBoost model. The [5] has emphasized studying the influence of vessel type on prediction, but the most significant characteristic discovered was MSTD, which was found after comparing trajectories using the SSPD approach. However, [17] utilized the ML model for the trajectory similarity and paired it with port frequencies to normalize the predictions. As a result, in this thesis, the challenge of predicting arrival port has been defined as a classification problem, and the best model discovered to tackle the problem is XGBoost.

RQ 1.b: What additional features can be added to improve the performance of the existing models?

As indicated in the preceding study questions' replies, most research is geographically confined. Some of the data aspects utilized in prior studies may be true for a limited area but not for data with a worldwide reach. Therefore, the studies that have been examined for consideration include [5, 17] since they are not geographically restricted. However, the research [5] has been taken into account since the same firm, Maritime Optima AS (MO), gives the data utilized for this thesis and the thesis presented in [5]. As a result, it was simple to compare the accuracy and feature importance with this study.

After researching the [5] article, it was discovered that the essential feature is MSTD, the least important is trajectory_length, and the prediction is also dependent on the kind of vessel which is defined by segment and sub-segment to which vessel belongs. To enhance the study, features such as season, probability, and distance_ratio were added while preserving all of the study's features. The probability feature appears to have high relevancy when predicting the arrival port; the distance_ratio is added as trajectory_length was the least important, so the distance_ratio can provide the same kind of information that is the position of the vessel, whether it is too close to departure port or have traveled a sufficient distance but in a better way, and it also appears to have high importance than the trajectory_length but not too high relative to other features. The season does not seem to be of great value, but experts believe that if the definition of seasons is changed, it may be of great importance.

Furthermore, the data in this research, as well as in [5], was inconsistent. The author of [5] used sampling approaches to ensure data consistency, which causes a lot of extra data addition or essential data removal to make the data consistent. The maritime professionals have also defended the omission of the arrival port in this thesis as the superior technique for keeping the data consistent. Only the ports that occur a few times have been removed. The model was trained independently on the various segments in this thesis, but the [5] was trained on the whole dataset and was used to assess the results of the model on other segments. The Table 6.1 compares the accuracy of the five segments considered for this thesis. And it can be seen that the model of this thesis performs better in terms of accuracy for all these segments.

Segment	**Accuracy in [5]**	**Proposed model accuracy**
Chemical	72.4%	73.95%
LNG	63.2%	73.29%
LPG	68.5%	77.45%
Dry_Bulk	67.1%	71.11%
Tanker	73.33%	75.03%

Table 6.1: Ablation study result for the five segments

6.2.2 RQ 2: What are the methods by which prediction of the availability of vessels at a port that carry specific cargo can be made?

Existing research did not give enough information to address this issue. As a result, there was an incentive to accomplish the tasks in this thesis. Thus, this thesis has

suggested a technique to achieve the problem. Furthermore, after developing the solution, the solution is presented to shipping professionals, who validate it and layout the commercial effect it would have on the marine sector. According to them, with few tweaks, the model will be incredibly advantageous and may be used in practice to estimate the availability of vessels at the port for a given cargo.

RQ 2.a: What kind of research methods have been used to predict the availability of vessels at a port for a specific cargo?

As previously stated, there was no prior work explicitly connected to the issue area. Some publications focused on the forecast of the arrival port or the prediction of the vessel's future location in time. Many research publications have focused on the Estimated Time of Arrival (ETA). Still, none of them have paired this with the prediction of arrival ports and then forecasting the ETA for those projected arrival ports exclusively. Instead, the papers have a specified port, or the two solutions were produced independently, and their correctness was tested separately. There were two articles were partly similar to the issue which were indicated in the Section 3.2, but neither focused on a particular cargo nor prediction for any time in the future.

RQ 2.b:If previous approaches have been limited, what approach could be used to predict the vessels at a port for a specific cargo?

As shown in Section 3.2, earlier techniques were limited in their ability to predict vessels as port. As a result, in this thesis, a technique for predicting the availability of a vessel at a port for a particular cargo was suggested. The procedure is divided into two steps: the first predicts the arrival ports for all ongoing vessels associated with the cargo, and the second calculates the ETA for those projected ports. The XGBoost model has been trained on different types of vessels' segments to forecast the arrival port. Two methods make the final prediction of the arrival port in this thesis: the ML model and the trajectory similarity technique SSPD. For calculating ETA, the collaborating company's routing engine is used, which was developed by shipping experts. It first determines the optimal route from the vessel's current position to the predicted port by avoiding all land parts and whether the vessel can pass through a small canal. If not, it takes the path that goes around the canal. Then, it calculates the ETA after constructing the route by calculating the route distance and the current vessel speed.

With the help of the result of the solution described in the thesis, suppose the broker wants to know the LNG vessels that may arrive in Rotterdam. The broker locates all active vessels, predicts their arrival port using the XGBoost model, trained on LNG data, and provides a final forecast based on trajectory similarity and the ML model. Then, send the projected arrival port and the current vessel location coordinates and speed to the route planner, who will give the ETA to the

predicted ports. And a query may be run from the final table to determine which vessels will arrive in Rotterdam in one week and their probability of reaching.

RQ 2.c: To what extent can this prediction be of practical value for the maritime industry?

The solution offered in the thesis has been discussed with shipping specialists to provide further insight into how the issue solution would be helpful. All shipping professionals play distinct roles in the maritime business. Therefore, the practical value is established from the perspective of the vessel owner, shipping analyst, and broker. The detailed description is presented in the Section 5.5.3; some of the essential points are described here.

According to the vessel owner, the solution will assist in correctly planning and scheduling the fleet of vessels, resulting in increased revenues as the fleet operates more efficiently. According to the broker, with the aid of the solution, more specific reports for the various vessels will be created, allowing the broker to provide better advice to customers who are vessel owners and cargo owners. According to the data analyst, it will aid in the systematic understanding of the supply situation in real-time, which will help in the prediction of the shipping market's ups and downs.

After analyzing the review of all the actors in the shipping business, there is optimism that the solution will be helpful if applied at the production level. However, the solution has some limitations, which are described below.

6.3 Limitations

In this part, there will be a discussion about the limitations and difficulties presented in the suggested solution to the thesis.

6.3.1 Voyage definition

MO's shipping specialists have defined the voyage concept utilized in this thesis. However, many incorrect voyages are still recorded, and many voyages are missed based on the specified criterion. For instance, if the voyage loses signal just outside the polygons established for the ports, the vessel arrival will not be logged. Still, the vessel will be departed from the same port in which it never arrived, resulting in an inconsistency in the vessel's travels. Furthermore, as the navigational statuses are utilized to define voyage arrival and departure, it has been observed that vessels change their statuses many times while within the port. As a result, many voyages arriving and departing at the same port are stored in the database.

The model presented in the thesis relies heavily on the current voyage database; thus, if the definition can be improved and in the future, more real voyages can be recorded and saved in the database, then the model's output will likely be enhanced. But the currently trained model will not work as with the new voyage definition; there will be a new dataset, so the model must be trained again. Therefore the present model has been limited to the current voyage definition only.

6.3.2 Season definition

To identify the seasons in the thesis, three months are grouped to define one season, resulting in four defined seasons. The voyages based on the departure timestamp have been distributed among these four seasons. However, when the importance of the feature is considered, the season has little impact on forecasting the arrival port. This can be because there is rarely any voyage that is three months long, so almost all the voyages will be completed in the same season. The definition of the seasons is thus constrained in the thesis. The seasons may be specified month by month, or months can be utilized as a feature in place of seasons to modify the season. Thus, in the future, a model can be defined with the new definition of the seasons to see the impact of the feature on the prediction of the arrival port.

6.3.3 Feature importance

From the Table 5.3 which defined feature importance for the prediction of arrival port by the ML model. It has been determined that the sspd_mstd feature, which is the MSTD derived from the SSPD technique of trajectory similarity, is of immense significance. All other characteristics have a low relevance percentage when predicting the arrival port. As the ML model is strongly reliant on the trajectory similarity value, if the MSTD prediction has an error, it may be anticipated that the ML model will similarly provide an incorrect result. Therefore this is a huge constraint in the thesis that the model is highly dependent on the trajectory similarity value rather than having equal importance for all the features.

6.3.4 Ports

As mentioned, the 'ports' table has been provided by the collaborating company of the thesis MO. The shipping expert for MO has identified those ports that have been utilized for trading and have a legitimate location as relevant in the 'ports' table. For the thesis, only relevant ports have been used. Some voyages in the dataset had irrelevant departure or arrival ports; those voyages were eliminated from the dataset during the training phase. However, the relevant ports vary with time; a port now labeled as irrelevant might be marked as relevant in the future. However, the trained ML model in the thesis has not seen the irrelevant ports

during training; therefore, it cannot make the prediction. Thus, the answer to the thesis is restricted to the present 'port' dataset.

6.3.5 Dataset imbalance

Its been seen in the thesis that the dataset is highly imbalanced; only some of the arrival ports account for most of the data, while the maximum amount of ports account for a very few percentages of data. Therefore, the arrival ports have been removed from the training purpose to make the data consistent. But in the future, it is possible that there can be voyages defined between the ports which have been removed. In that case, it will be possible to add them to the dataset as they have a relevant percentage of data required to be predicted. In that case, the model needs to be trained again with the new dataset, which might include ports that have been removed before. Therefore, the model specified in the thesis must be trained according to time progression to include the new ports that now have adequate voyage data to be predicted by the ML model.

6.4 Conclusion and future work

The shipping industry is complex and constantly affected by supply and demand for the cargo. Keeping track of the fluctuation of the shipping industry is essential to being a successful shipowner or a broker. The availability of a vessel at a port for a specific cargo can affect the shipping rates and the cargo's movement. By predicting the availability of a vessel, it is possible to factor in this volatility and make informed decisions about shipping the cargo. Therefore, in this thesis, a technique has been described that may assist in estimating the availability of a vessel at the port for specific cargo and the possibilities of the vessel arriving to be sure to what extent the prediction is accurate. The significance of this thesis has been validated by the shipping specialists who operate in different areas of the shipping industry. According to all of them, it will be of great significance for the shipping business.

The thesis provided a solution for the prediction of the arrival port, which is highly dependent on the previous trajectory. But in the future, it will be interesting to have a solution that is not dependent on the previous trajectory. In that way, there is a possibility of predicting the arrival port and the next port, also at which the vessel will go from the predicted port.

The solution for predicting the availability of vessels has been defined in the two steps. It will be of high importance in the future if only one model can predict the availability of vessels. So as input, the features related to the port will be given, and the model will predict the vessels that can arrive at the port.

The model may be modified segment by segment, concentrating exclusively on big vessels. Because according to the experts interviewed, there would be more excellent economic value if the model could be trained on segments with a concentration exclusively on big vessels to increase model accuracy and then estimate vessel availability at the port.

The seasons are not a significant feature in this thesis, yet cargo movement is heavily dependent on seasonality. So, the model should be tested with other season definitions in the future. For example, instead of a season, a month feature may be included, which means that instead of grouping months to define one season, there will be twelve months depending on departure timestamp as a feature. It would be interesting to learn how months affect arrival port prediction.

6.5 Concluding remark

At the outset of the thesis, there is no publicly known systematic method for obtaining information on all vessels. The shipbroker gathers information from numerous sources and reports it to the cargo owners and vessel owners. However, there is no transparency in the system; it is based on information from different members of the shipping sector, which may be inaccurate as well. The technique in the thesis solves the issue to some degree by forecasting the availability of the vessels at the port. Along with the prediction, the thesis provides a probability factor to show to what degree the expected answer may be believed. Finally, the thesis emphasizes the actual use of the vessel availability solution at a port that the working shipping experts provide. As a result, the solution presented in the thesis will be effective when applied at the production level and helpful for the marine industry. While this thesis expands on the prior work of the thesis [5], it broadens the viewpoint beyond port prediction to include the computation of ETA, which results in the forecast of vessel availability at a port. The thesis demonstrates a direct approach to addressing this, which yielded promising results; however, future work should improve the results presented and investigate other predictive methods that address the problem of vessel availability prediction as one of the significant challenges in maritime logistics.

Bibliography

[1] T. Chen and C. Guestrin, 'XGBoost,' in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. [Online]. Available: https://doi.org/10.1145%2F2939672.2939785.

[2] *Difference between a shipbroker and ship charterer*, https://www.shippingandfreightresource.com/shipbroker-and-ship-charterer/, (Accessed on 05/27/2022).

[3] S. Skallist, 'Which behaviours do shipbrokers use to create interpersonal trust and relationships with clients?' M.S. thesis, University of South-Eastern Norway, 2018.

[4] *Cargo shipping market revenue & size | global forecast [2028]*, https://www.fortunebusinessinsights.com/cargo-shipping-market-102045, (Accessed on 05/28/2022).

[5] M. Omholt-Jensen, *Vessel destination forecasting based on historical ais data*, eng, 2021. [Online]. Available: https://hdl.handle.net/11250/2778076.

[6] L. Wu, Y. Xu and F. Wang, 'Identifying port calls of ships by uncertain reasoning with trajectory data,' *ISPRS International Journal of Geo-Information*, vol. 9, no. 12, 2020, ISSN: 2220-9964. DOI: 10.3390/ijgi9120756. [Online]. Available: https://www.mdpi.com/2220-9964/9/12/756.

[7] T. Mestl, D. Gl, H. Norway and K. Dausendschön, 'Port eta prediction based on ais data,' May 2016.

[8] [Online]. Available: http://psimpl.sourceforge.net/douglas-peucker.html.

[9] Y. Suo, W. Chen, C. Claramunt and S. Yang, 'A ship trajectory prediction framework based on a recurrent neural network,' *Sensors*, vol. 20, no. 18, 2020, ISSN: 1424-8220. DOI: 10.3390/s20185133. [Online]. Available: https://www.mdpi.com/1424-8220/20/18/5133.

[10] D. Demyen and M. Buro, 'Efficient triangulation-based pathfinding,' in *Aaai*, vol. 6, 2006, pp. 942–947.

[11] Q. Liu and Y. Wu, 'Supervised learning,' Jan. 2012. DOI: 10.1007/978-1-4419-1428-6_451.

[12] J. J. Oliver, R. A. Baxter and C. S. Wallace, 'Unsupervised learning using mml,' in *ICML*, Citeseer, 1996, pp. 364–372.

[13] V. Heidrich-Meisner, M. Lauer, C. Igel and M. Riedmiller, 'Reinforcement learning in a nutshell,' Jan. 2007, pp. 277–288.

[14] D.-D. Nguyen, C. Le Van and M. I. Ali, 'Vessel trajectory prediction using sequence-to-sequence models over spatial grid,' in *Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems*, ser. DEBS '18, Hamilton, New Zealand: Association for Computing Machinery, 2018, pp. 258–261, ISBN: 9781450357821. DOI: 10.1145/3210284.3219775. [Online]. Available: https://doi.org/10.1145/3210284.3219775.

[15] R. Gouareb, F. Can, S. Ferdowsi and D. Teodoro, 'Vessel destination prediction using a graph-based machine learning model,' in *Network Science*, P. Ribeiro, F. Silva, J. F. Mendes and R. Laureano, Eds., Cham: Springer International Publishing, 2022, pp. 80–93, ISBN: 978-3-030-97240-0.

[16] B. B. Magnussen, N. Bläser, R. M. Jensen and K. Ylänen, 'Destination prediction of oil tankers using graph abstractions and recurrent neural networks,' in *Computational Logistics*, M. Mes, E. Lalla-Ruiz and S. Voß, Eds., Cham: Springer International Publishing, 2021, pp. 51–65, ISBN: 978-3-030-87672-2.

[17] 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.

[18] C.-X. Lin, T.-W. Huang, G. Guo and M. D. F. Wong, 'Mtdetector: A high-performance marine traffic detector at stream scale,' in *Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems*, ser. DEBS '18, Hamilton, New Zealand: Association for Computing Machinery, 2018, pp. 205–208, ISBN: 9781450357821. DOI: 10.1145/3210284.3220504. [Online]. Available: https://doi.org/10.1145/3210284.3220504.

[19] M. Bachar, G. Elimelech, I. Gat, G. Sobol, N. Rivetti and A. Gal, 'Venilia, on-line learning and prediction of vessel destination,' in *Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems*, ser. DEBS '18, Hamilton, New Zealand: Association for Computing Machinery, 2018, pp. 209–212, ISBN: 9781450357821. DOI: 10.1145/3210284.3220505. [Online]. Available: https://doi.org/10.1145/3210284.3220505.

[20] V. Roşca, E. Onica, P. Diac and C. Amariei, 'Predicting destinations by nearest neighbor search on training vessel routes,' in *Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems*, ser. DEBS '18, Hamilton, New Zealand: Association for Computing Machinery, 2018, pp. 224–225, ISBN: 9781450357821. DOI: 10.1145/3210284.3220509. [Online]. Available: https://doi.org/10.1145/3210284.3220509.

[21] G. B. Karataş, P. Karagoz and O. Ayran, 'Trajectory prediction for maritime vessels using ais data,' in *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, ser. MEDES '20, Virtual Event, United Arab Emirates: Association for Computing Machinery, 2020, pp. 48–54, ISBN: 9781450381154. DOI: 10.1145/3415958.3433079. [Online]. Available: https://doi.org/10.1145/3415958.3433079.

[22] A. Dobrkovic, M.-E. Iacob and J. van Hillegersberg, 'Maritime pattern extraction and route reconstruction from incomplete ais data,' *International journal of Data science and Analytics*, vol. 5, no. 2, pp. 111–136, 2018.

[23] 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*, ACM, Jun. 2018. DOI: 10.1145/3210284.3220502. [Online]. Available: https://doi.org/10.1145%2F3210284.3220502.

[24] H. Jung, K.-W. Lee, J.-H. Choi and E.-S. Cho, 'Bayesian estimation of vessel destination and arrival times,' in *Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems*, ser. DEBS '18, Hamilton, New Zealand: Association for Computing Machinery, 2018, pp. 195–197, ISBN: 9781450357821. DOI: 10.1145/3210284.3220501. [Online]. Available: https://doi.org/10.1145/3210284.3220501.

[25] C. K. Pham, *Predicting the next port visit of a vessel using ais data*, 2019-12. [Online]. Available: http://hdl.handle.net/10945/64046.

[26] T. Wang, C. Ye, H. Zhou, M. Ou and B. Cheng, 'Ais ship trajectory clustering based on convolutional auto-encoder,' in *Intelligent Systems and Applications*, K. Arai, S. Kapoor and R. Bhatia, Eds., Cham: Springer International Publishing, 2021, pp. 529–546, ISBN: 978-3-030-55187-2.

[27] A. Dobrkovic, M.-E. Iacob and J. van Hillegersberg, 'Using machine learning for unsupervised maritime waypoint discovery from streaming ais data,' in *Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business*, ser. i-KNOW '15, Graz, Austria: Association for Computing Machinery, 2015, ISBN: 9781450337212. DOI: 10.1145/2809563.2809573. [Online]. Available: https://doi.org/10.1145/2809563.2809573.

[28] W. Wang, C. Zhang, F. Guillaume, R. Halldearn, T. S. Kristensen and Z. Liu, 'From ais data to vessel destination through prediction with machine learning techniques,' *Artificial Intelligence: Models, Algorithms and Applications*, p. 1, 2021.

[29] D. Alizadeh, A. A. Alesheikh and M. Sharif, 'Prediction of vessels locations and maritime traffic using similarity measurement of trajectory,' *Annals of GIS*, vol. 27, no. 2, pp. 151–162, 2021. DOI: 10.1080/19475683.2020.1840434. eprint: https://doi.org/10.1080/19475683.2020.1840434. [Online]. Available: https://doi.org/10.1080/19475683.2020.1840434.

[30] L. Ming and J.-x. Liu, 'Prediction of the amount of vessels arriving at inland port based on time series analysis,' in *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, 2017, pp. 831–835. DOI: 10.1109/ICTIS.2017.8047864.

[31] D. Yang, L. Wu and S. Wang, 'Can we trust the ais destination port information for bulk ships?–implications for shipping policy and practice,' *Transportation Research Part E: Logistics and Transportation Review*, vol. 149, p. 102 308, 2021, ISSN: 1366-5545. DOI: https://doi.org/10.1016/j.tre.2021.102308. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136655452100082X.

Appendix A

Additional Material

The code for connecting the routing engine of Maritime Optima AS (MO) to calculate the ETA for the predicted ports. The code takes input the excel file which contains current position coordinates, predicted port coordinates and the vessel current speed. It gives out also an excel file which contains Estimated Time of Arrival (ETA) in hours.

Code listing A.1: Go code used to calculate ETA

```
package main

import (
        "context"
        "flag"
        "fmt"
        "io/ioutil"
        "os"
        "time"

        "github.com/MaritimeOptima/services/pkg/geometry/s2util"
        "github.com/MaritimeOptima/services/routing-engine-v2/pkg/rev2client"
        "github.com/golang/geo/s2"
        "github.com/sirupsen/logrus"

        "github.com/gocarina/gocsv"
)

type Voyage struct {
        IMO      int64 `csv:"imo"`
        VoyageID int64 `csv:"voyage_id"`

        StartLat float64 `csv:"start_lat"`
        StartLon float64 `csv:"start_lon"`

        EndLat float64 `csv:"end_lat"`
        EndLon float64 `csv:"end_lon"`

        ExpectedDurationHours *float64 `csv:"expected_duration_hours"`
        Speed float64 `csv:"imo"`

        // EstimatedDuration *float64

```

```
34          // LOA   float64
35          // Draft float64
36          // Beam  float64
37  }
38
39  func readCSV(path string) ([]*Voyage, error) {
40          file, err := os.OpenFile(path, os.O_RDWR|os.O_CREATE, os.ModePerm)
41          if err != nil {
42                  return nil, err
43          }
44          defer file.Close()
45
46          voyages := []*Voyage{}
47
48          if err := gocsv.UnmarshalFile(file, &voyages); err != nil {
49          // Load clients from file
50                  return nil, err
51          }
52
53          return voyages, nil
54  }
55
56  func main() {
57          csvFile := parseFlags()
58
59          cli, err := rev2client.NewClient("routing-engine-v2.services:3000")
60          if err != nil {
61                  err.Fatal(logrus.StandardLogger())
62          }
63          defer cli.Close()
64
65          voyages, eerr := readCSV(csvFile)
66          if err != nil {
67                  logrus.WithError(eerr).Fatal("failed to read csv")
68          }
69
70          invalids := make(map[string]*Voyage)
71
72          for _, voyage := range voyages {
73                  timer := time.Now()
74
75                  route, err := cli.GetRoute(context.Background(), rev2client.RouteRequest{
76                          Legs: []rev2client.LegRequest{
77                                  {
78                                          From: rev2client.Waypoint{
79                                                  Point: pointFromLatLon(voyage.StartLat, voyage.
                                                        StartLon),
80                                          },
81                                          To: rev2client.Waypoint{
82                                                  Point: pointFromLatLon(voyage.EndLat, voyage.EndLon)
                                                        ,
83                                          },
84                                  },
85                          },
86                  })
87                  if err != nil {
88                          // err.Fatal(logrus.StandardLogger())
89                          logrus.Warn("Invalid end or start position")
90                          logrus.Warnf("Voyage: %+v", voyage)
91                          logrus.Warn(err)
```

```
92
93                      invalids[fmt.Sprintf("%f,%f", voyage.EndLat, voyage.EndLon)] = voyage
94                      continue
95              }
96
97              polyLine := route.MainRoute.Legs[0].Polyline
98
99              distKM := s2util.AngleToKm(polyLine.Length())
100             durationHours := float64(distKM) / Speed // 14 knots ca
101
102             logrus.WithFields(logrus.Fields{
103                     "calculation_time": time.Since(timer).Seconds(),
104                     "dist_km":          distKM,
105                     "duration_hours":   durationHours,
106             }).Info("found route")
107
108             voyage.ExpectedDurationHours = &durationHours
109     }
110
111     logrus.Warnf("%d invalid routes", len(invalids))
112
113     csvOutput, eerr := gocsv.MarshalString(voyages)
114     if eerr != nil {
115             logrus.WithError(eerr).Fatal("failed to marshal csv")
116     }
117
118     eerr = ioutil.WriteFile("./output.csv", []byte(csvOutput), 0644)
119     if eerr != nil {
120             logrus.WithError(eerr).Fatal("failed to write csv")
121     }
122 }
123
124 func parseFlags() string {
125     var csvFile string
126
127     flag.StringVar(&csvFile, "csvFile", "./voyages.csv", "Voyage CSV file")
128
129     flag.Parse()
130     if csvFile == "" {
131             fmt.Fprintf(os.Stderr, "Usage of %s:\n", os.Args[0])
132             flag.PrintDefaults()
133             os.Exit(1)
134     }
135
136     return csvFile
137 }
138
139 func pointFromLatLon(lat, lng float64) s2.Point {
140     return s2.PointFromLatLng(s2.LatLngFromDegrees(lat, lng))
141 }
```

NTNU
Norwegian University of
Science and Technology

MARITIME OPTIMA