



UNIVERSITY OF
LIVERPOOL

Management School

The University of Liverpool

Business Analytics and Big Data MSc Project

ETA Prediction for Container Ports using Machine
Learning Techniques at Port of Southampton

Project No: SONG-01

Final Report

by

Masood Salman Choudhury

Student ID: 201505340

Date: 21/11/21

Supervisor: Dr Dongping Song

Executive Summary

90% of global trade is carried by water, with container vessels accounting for the majority of it. However, uncertainty regarding the estimated arrival time of container vessels has a negative impact on the planning activities of stakeholders involved in container shipping.

This project aims to provide stakeholders in the Port of Southampton with a vessel ETA prediction tool in order to increase their operational efficiency. To begin, this project establishes the value and impact that an ETA prediction tool has on various stakeholders. Then, based on previous research, the factors affecting the vessel's ETA are identified. To define the fundamental data requirements for building a machine learning model. Once the factors were identified, the performance of four machine learning models was compared (Decision Tree, Random Forest, Support Vector Machine, Neural Network). Furthermore, the machine learning model's feature importance was extracted to analyze and establish new relationships, including covid cases and vessel delay.

After developing models and evaluating the results, the practical applications and implementation of the model involving stakeholder were discussed.

Acknowledgement

First and foremost, I am incredibly grateful to my supervisor, Prof. Dr Dongping Song, for his continuous support and patience. Without his guidance, this project would not be possible. Also, I would like to thank Dr Akshit Singh & Dr Jorge Hernandez Hormazabal for providing all the necessary guidelines for completing this project.

Massive thanks to my colleagues in this project Mr Aimal Khan and Miss Qian Lan, for their support. I would also like to thank my parents and my brother for supporting and believing in me. Last but not least, I am grateful to all my friends who were a source of inspiration throughout the project.

Contents

Executive Summary	2
Chapter 1 - Introduction	10
1.1 Background	10
1.2 Study Area	11
1.3 Significance	12
1.3.1 Practical Significance	12
1.3.2 Scientific Significance	14
1.4 Aim and Objectives	15
1.5 Research Methods	15
1.6 Structure of the Report	17
Chapter 2 - Literature Review	18
2.1 Importance of ETA Predictions	18
2.1.1 Vessels Operators	19
2.1.2 Terminal and Seaport Operator	20
2.1.3 Hinterland Transportation	22
2.1.4 Importers	23
2.2 Current Studies on ETA prediction	23
2.3 Factors Affecting ETA	25
2.4 Research Gap	27
Chapter 3 - Methodology	29
3.1 CRISP-DM Process Model	29
3.2 Business Understanding	31
3.3 Data Understanding	32
3.4 Data Preparation	35

3.4.1	Correcting Data format	35
3.4.2	Feature Engineering	36
3.4.3	Data Normalization	38
3.4.4	Categorizing Target Variable	39
3.4.5	Encoding Data	40
3.5	Modeling	42
3.5.1	Decision Tree	43
3.5.2	Random Forest	46
3.5.3	Support Vector Machine	48
3.5.4	Neural Network	49
3.6	Evaluation	52
Chapter 4 - Analysis and Findings		55
4.1	Initial Data Exploration	55
4.2	Model with All Features	60
4.3	Feature Selection	62
4.3.1	Result After Feature Selection	63
4.3.2	Feature Importance	64
4.3.3	Discussion	67
4.4	In depth Class Label based Analysis of RF Model	68
4.5	RF Model with 2 Class Labels	69
Chapter 5 - Practical Implementation		74
5.1	Effect on Stakeholder	74
5.2	Implementing at Port of Southampton	76
Chapter 6 - Conclusion		78
6.1	Main Findings	78
6.2	Limitation	80

6.3 Areas for further Research	80
Reference List	82
Appendix	89

List of Figures

Figure 2.1 landside and Seaside Operations	21
Figure 3.1 - CRISP-DM Process Phase	30
Figure 3.2 One-Hot Encoding	41
Figure 3.3 - Multilayer Neural Network	50
Figure 4.1 - Correlation Matrix of Size and other volumetric metrics with Delay variable	55
Figure 4.2 - Regression Line, x = Built Year of Vessel, y = Size of Vessel in TEU	56
Figure 4.3 - Boxplot, x = Berth, y = Size of Vessel in TEU	56
Figure 4.4 - Correlation Matrix - Correlation between Delay and Continuous variable	57
Figure 4.5 - Distribution plot of Days Before ETA Info	58
Figure 4.6 - Regression line of Delay and Days Before ETA info	58
Figure 4.7 - Regression line of Delay and weekly new Covid cases UK	59
Figure 4.8 - Class labels distribution	60
Figure 4.9 - Accuracy of Models - All Features - Boxplot	61
Figure 4.10 - Macro F1 of Models - All Features - Boxplot	61
Figure 4.11 - RF Models Feature Importance - All features	62
Figure 4.12 - Accuracy of Models - Selected Features - Boxplot	63
Figure 4.13 - Macro F1 of Models - Selected - Boxplot	63
Figure 4.14 - RF Models Feature Importance - Selected Features	64
Figure 4.15 - NextPort and Service - Scatter Plot	66
Figure 4.16 - Two Classes label Distribution	71
Figure 4.17 - Feature importance of 2 Class Label RF model	72
Figure 4.18 - Test Accuracy with Number of trees in RF model	73

List of Tables

Table 2.1 - Current Studies of ETA prediction: Summarised	25
Table 2.2 - Factor affecting Vessel ETA with Reference	26
Table 3.1 - DPWorld Southampton Data	33
Table 3.2 - Clarksons World Fleet Register Data	34
Table 3.3 - UK Covid Data	34
Table 3.4 - Raw format of DP World Southampton Data	35
Table 3.5 - After fixing the formatting irregularities	36
Table 3.6 - Same Vessel Name multiple phases with different Ref	37
Table 3.7 - INBOUND and ARRIVED for each unique Vessel	37
Table 3.8 - Table with ETA,ATA and Delay	38
Table 3.9 - Label Encoded Data set features	42
Table 4.1 - Mean accuracy and F1 score of Model	61
Table 4.2 - Mean accuracy and F1 score of Models	63
Table 4.3 - Model Comparison with Nextport and Service Feature	66
Table 4.4 - Training Accuracy and F1 score of RF Model	68
Table 4.5 - Confusion Matrix for Multi-Class Classification	68
Table 4.6 - Class based performance metrics of Test Data	69
Table 4.7 - Training Accuracy and F1 score of RF Model with 2 Class Labels . .	70
Table 4.8 - 2 Classes performance metrics on Test Data	70

List of Abbreviations

Abbreviations	Meaning
NN	Neural Network
MLP	Multilayer Perceptron
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machine
Dwt	Deadweight Tonnage
GT	Gross Tonnage
TEU	Twenty-foot Equivalent Unit
AGV	Automated Guided Vehicles
ETA	Estimated Time of Arrival
ETD	Estimated Time of Departure
ATA	Actual Time of Arrival
CRISP-DM	Cross-Industry Standard Process for Data Mining
COVID-19	Coronavirus Disease 2019
KNN	K-Nearest Neighbors
AIS	Automatic Identification System
TAT	Turnaround Time
RBF	Radial basis function
ReLU	Rectified Linear activation function

Chapter 1 - Introduction

*This **Chapter** introduces the importance of seaborne trade and how uncertainty in vessel arrival can affect stakeholders of containership transport. **Section 1.1** Emphasizes the importance and scope of sea trade. **Section 1.2** Highlights the Port of Southampton and its relevance. **Section 1.3** Explores the effect and significance of vessel uncertainty on stakeholders and scientific research. **Section 1.4** Describe the aim and objective of the project. **Section 1.5** summarizes the research approach and methods. Final **Section 1.6** explains the overall structure of the whole project.*

1.1 Background

As a result of today's globalized economy, items of all types are created and shipped around the globe. More than 90% of this trade is carried out by sea because of its unparalleled capacity, ability to travel long distances economically, and environmental friendliness (Meijer, 2017; OECD, 2019). All kinds of items are transported via maritime routes that traverse continents through seaports, such as raw materials, parts, and finished goods. Thus, making seaports the backbone of maritime transport and the modern economy. International supply networks rely on seaports to a large extent. Many studies are emphasizing the importance of ports and container terminals in the supply chain and global trade (Robinson, 2006; Wang and Cullinane, 2006). Unloading, loading containers from vessels, warehousing, and linking to inland logistics are a few key responsibilities of seaports, which are critical nodes in the supply chain process (Montwiłł, 2014). Despite the fact that there are thousands of ports all over the world, the top 20 busiest ports handle over half of all container throughput globally. Thus, making port efficiency is a prominent topic among researchers. Port operations for containers can be divided into two groups: landside operations and seaside operations. Landside deals with the gate, the yard, and the quay. Multiple comprehensive studies that cover a wide range of landside

operations management challenges such as quay crane and yard crane scheduling, storage management, inter-terminal truck/AGV routing, and terminal gate appointment issues. Seaside consist of the berth, the channel, and the anchorage. This also comes with its own set of issues, including berth allocation and channel and anchorage utilization (Wei et al., 2020). The majority of container port operations required advanced planning to effectively allocate resources such as berth allocation, crane scheduling, warehousing, and logistics for hinterland transport for upcoming arriving vessels. Thus, it is crucial to know the vessel’s arrival time to better prepare for it. However, the unpredictability of container vessel arrival times at the container ports is one of the major issues (Gómez, Camarero and Molina, 2016). Despite contractual obligations to report the estimated arrival time (ETA) 24 hours in advance, ship operators often have to adjust it because of unexpected occurrences such as weather conditions, delays in a previous port, and etc. (Fancello et al., 2011). A vessel delay can lead to delayed docking, problem in berth allocation, and crane schedules, leading to more delay for hinterland transportation. For planners, decision-making procedures relating to port actives scheduling and resource allocation might be quite complex at times without the assistance of appropriate methodological tools for accurately predicting ETA.

The purpose of this project is to provide a method to predict the estimated arrival time of containerships reliably and to analyze the utility of such a decision support system for port operators and other stakeholders involved in container transportation.

1.2 Study Area

This study will focus on the Southampton seaport as a case study mainly due to the availability of data. Southampton is the second-largest container port in the UK, known as Britain’s Gateway to the World, with its ideal location on the south coast adjacent to key shipping lanes connecting the UK to European and global markets (DP World, 2021). The port is the UK’s most important export port and a vital link in supply chains for

businesses and industries across the country. Its state of art five berths container port that DP World operates can handle over 1.9 million TEUs annually with its 500m of the quay and 16.5m water depth that can handle even the largest container vessels, thus Making it the second largest and busiest port in the UK. It employs 45,600 people and generates £2.5 billion in revenue for the UK economy each year (Associated British Ports, 2019).

1.3 Significance

1.3.1 Practical Significance

Over 80% of global merchandise is delivered by sea and managed by ports. Containers transport almost 70% of the value of this seaborne trade (UNCTAD, 2020). These containers are processed at seaport terminals, which serve as critical interfaces between land- and sea-based transportation and between multiple modes of transport. Thus, seaports have been designated as crucial infrastructures, which are critical components that affect a country's economic and social well-being (Knowles, Shaw and Docherty, 2008). This means that ensuring deep seaports' smooth and efficient operation is critical for the efficient and rapid distribution of commodities to their final inland destinations. As container transport has other stakeholders besides Seaports. Such as Terminal Operators, Vessels Operators, Hinterland Transportation Partners and Importers.

Terminal and Seaport Operator

Seaport Operators or Seaport authorities are responsible for the management of seaports in general. A seaport is a geographic area where ships are brought to unload and load cargo. Seaports usually include multiple terminals which are specialized in different cargo, i.e. (Container, Dry Bulk, Oil). Seaport operators aim to improve the overall efficiency of the seaport, which include providing different terminals for different cargo, Storage for storing cargo, also integrating hinterland transport for efficient transportation of goods. Whereas the Terminal Operator is responsible for managing all the operations that take

place in the terminal, these can be categorized into Seaside and Landside operations. This operation can range from assigning berth, unloading cargo ships, workforce management, and many more. This will be elaborated on in Chapter 2 (Literature review section). These Operations are complex and require strategic planning in advance to efficiently manage and allocate seaport resources. However, Uncertainty in vessel arrival can hamper the planning process done by terminal operators, which can lead to inefficient allocation of resources (Wei et al., 2020). Also, as terminals are part of a seaport, thus, affecting the efficiency of a seaport in general.

Vessels Operators

Shippers employ vessels to transport goods between an origin and a deep-sea port destination. Typically, shipping corporations manage a fleet of vessels in order to grab the maximum amount of market feasible and maximize their profit. Additionally, in order to maximize profit per voyage, vessels run at the lowest possible speed to conserve fuel while still arriving on time, due to the fact that fuel accounts for more than half of the entire expense associated with a voyage (Parolas, 2016). An inaccurate ETA might result in mismanagement of the vessel's speed, resulting in increased fuel expenses or a delay in reaching the target (Stopford, 2009).

Hinterland Transportation

Hinterland transport is the transportation network that exists on the landward side of a seaport and is used to transfer cargo to and from the seaport. The planning operations of partners in hinterland transport are reliant upon the availability of precise ETA information for sea vessels. Additionally, the Port of Southampton (DP World) operator operates its own Hinterland transit service in conjunction with other third parties companies. As all hinterland transportation activities are dependent on the vessel arriving at the seaport on time, any delay will disrupt the plan and flow of the hinterland supply chain, resulting

in cost increases and further delays.

Importers

Importers are the purchasers and receivers of the products contained within the container-ship. Any delay in the overall supply chain process can result in unsatisfied importers, particularly if the goods are perishable. If this is the case, a loss may result. As demonstrated in the preceding section, uncertainty over vessel arrival can result in a variety of delays throughout the subsequent supply chain process.

Therefore, it is quite evident that precisely forecasting the estimated arrival time for sea vessels at a port is critical for the cost-effective execution of port operations and the activities of stakeholders in the overall container transport supply chain.

1.3.2 Scientific Significance

In the recent era, there has been increased interest and research of ETA prediction on container vessels. The most notable works were done by (Parolas, 2016) and (Meijer, 2017). Both authors focused on Port of Rotterdam and Machine learning models for prediction. Parolas thesis focused on ETA predicting container vessels using weather and AIS Data and was focused on a specific route, while (Meijer, 2017) research extended the research done by Parolas by separating predicting model based on the route the vessel was using.

Even with increased research in ETA prediction of container vessels, no study focuses on the UK's port, specifically Southampton, as per the knowledge of the writer of this paper. This study will focus on predicting the arrival of container vessels without using any weather data, unlike the previously mentioned researches. This study will focus on predicting arrival time based on port call data and vessel characteristics that are publicly available and easy to collect. Moreover, this study also takes effect of the covid pandemic

in regard to vessel uncertainty.

1.4 Aim and Objectives

The project’s primary research goal is to pre-process the data from a variety of sources and build effective predictive models to anticipate the estimated time of arrival (ETA) of containerships at the Port of Southampton by using various machine learning models to aid in the planning activities of stakeholders involved in container transport in terms of cost and efficiency.

The specific objectives are stated as research questions as follows:

R01 - What benefit will precise forecast of containership arrival times at container terminals provide for port planning authorities and other stakeholders involved in container transport?

R02 - What are the primary factors influencing vessel delays?

R03 - Which machine learning algorithm produces the most accurate results?

R04 - Is Vessel ETA affected by covid cases?

1.5 Research Methods

Two different research approaches will be taken depending on the objectives to answer the paper’s research aim and objectives.

To answer the study’s primary objective, ‘Predicting ETA on container vessels at Southampton port’. The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) process model will be followed to create the machine learning models, which will be further elaborated in the Methodology section (Chapter 3). For any machine learning model, data is the foremost requirement. Therefore, data such as historical port call data collected by DP-World Southampton and Vessel characteristics data from Clarksons World Fleet Reg-

ister will be utilized for the prediction model. Historical port call data include data from 4 July 2020 to 5 Oct 2021 of Southampton port. Following the data mining procedure, data cleaning and pre-processing are required to develop the input features that will be utilized as a predictor to estimate the arrival time of containerships. Moreover, to establish the relation between Covid Pandemic and ETA, daily new cases data from a UK government repository for Covid Cases will be used.

A quick summary of features of the data obtained.

Features from DP-World Southampton data:

Vessel Name, Service, Phase, ETA, Berth, Side To, ETD, Next Port of Call, Ref.

Features from Clarksons World Fleet Register data:

Vessel Size, Dwt, GT, Flag, Built Date, Builder, Operator, Engine Type.

Features of UK Covid Cases:

Date, New Cases.

All features will be further explained in Section 3.3 Data Understanding. The expected time of arrival will be estimated based on the aforementioned inputs. Different machine learning algorithms such as Decision Tree, Random Forest, Neural network (Multilayer perceptron), and support vector machines will be employed. All of these machine learning models are based on supervised learning; that is, they learn from historical data to predict future outcomes. Why these specific machine learning models were selected will be discussed in Section 3.5 Modeling.

Along with achieving the primary objective of vessel ETA prediction with these machine learning models, these models will help answer sub-objectives R02, R03 and R04. As we

can analyze the created machine learning model and explore which factors affect the outcome more, and to tackle the R03, we will be comparing various machine learning models. For R04, the Model will also reveal if there is any relation between Vessel ETA and covid cases. Methods for obtaining the result will be further elaborated in Section 3.5 Modeling.

Moreover, to answer the R01 ‘Benefit of accurate ETA prediction for stakeholders. A Literature Review in Section 2.1 will be conducted to identify how an ETA prediction tool will affect stakeholders involved in container vessel transportation. ScienceDirect and Google Scholar databases will be used to collect the relevant articles and journals regarding it.

1.6 Structure of the Report

The report is structured as follows: **Chapter 1** Highlight the problem and importance of uncertainty of vessel arrivals at the Port of Southampton and the project’s research objective. **Chapter 2** will provide an overview of the literature on the topic to understand current forecasting techniques and assess the impact that better ETA predictions can have on the Port operators and stakeholders. Then, **Chapter 3** explains the methodology followed for predicting the Estimated time of arrival for vessels. **Chapter 4** analyses and presents the results obtained using the proposed Machine learning models. **Chapter 5** discusses the practical implications of result. Finally, the report is concluded in **Chapter 6** by summarizing the main findings of the project ,limitation and offering areas for further research.

Chapter 2 - Literature Review

*This **Chapter** aims to review relevant literature and identify the research gap. **Section 2.1** addresses the **RO1** of the thesis by emphasizing the importance of ETA prediction to its various stakeholders. **Section 2.2** summarises the existing approaches to ETA prediction based on literature reviews. **Section 2.3** identifies numerous factors affecting the vessel's ETA. **Section 2.4** dives into the Research gap of existing literature and how this research differentiate from those.*

2.1 Importance of ETA Predictions

This section elaborates the Importance ETA prediction will have on the parties and stakeholders involved in the Container transport industry while also answering R01, i.e. “What benefit will a precise forecast of containership arrival times at container terminals provide for port planning authorities and other stakeholders involved in container transport?”

Transportation using containers is one of the most important modes of transportation in the transportation industry. The majority of containers are transported by container ships that cruise around the world, visiting various ports. A lot may happen during these trips, which can alter the scheduled travel times of these vessels. Although as per the contractual agreement, the ship operator must announce ETA to the destined port 24 hours in advance (Fancello et al., 2011). However, these arrival times are approximated by a vessel's captain and are rarely precise, which profoundly impacts container terminal procedures and increases supply chain costs (Parolas, 2016).

The accuracy of the arrival of the container vessels determines every step in the supply chain. To be efficient, a port and hinterland network must receive reliable information regarding the arrival time of cargo aboard a container vessel. If a forecast is inaccurate,

the negative consequences affect the whole supply chain and exacerbate at each link in the supply chain; this is known as the bullwhip effect (Lee, Padmanabhan and Whang, 1997). According to a Drewry Shipping Consultants report, more than 40% of vessels operating on global liner lines had been delayed for one or more days (Drewry, 2006). Also, as per the "Sea-Intelligence Global Liner Performance August 2018 Report," the average delay of vessels is around 3.5 days. This can cause planning issues for stakeholders involved in container shipping. However, issues can be minimized if accurate information is exchanged and shared across the supply chain. Numerous stakeholders are engaged in the container transport process, beginning from the moment a business requires the transportation of container goods from its initial place to its final destination. The following stakeholders of Port of Southampton will benefit from the proposed ETA Prediction tool:

- Vessel Operators.
- Terminal and Seaport Operator.
- Hinterland Transportation Partners
- Importers

2.1.1 Vessels Operators

Shippers hire vessels to move their cargo between an origin and a destination deep-sea port. Shippers do not usually hire the entire vessel with just enough space or slot there to transport their goods. Typically, shipping corporations operate a fleet of vessels specialized in delivering specific types of cargo, such as dry cargo or liquid cargo. The shipping sector is extremely competitive, with numerous corporations engaged. As a result, the sea shipping business operates nearly entirely according to the norms of perfect competition, with prices substantially influenced by supply and demand (Stopford, 2009). When the number of ships available for cargo transportation is limited, carriers can charge high fees and maintain a high-profit margin due to increased demand. However, when more ships are

built to take advantage of the potential for high prices, which leads the market becomes saturated, with an excess of ships and insufficient goods to transport. The shipping industry then begins a phase of decline. During this period, cost reduction becomes critical to maintaining profitability. In such instances, it is common practice to operate the vessels at the lowest possible speeds to save on fuel consumption, as fuel consumption accounts for most of a ship’s expenses, accounting for over 60% of the total cost (Tran and Lam, 2021). So, a precise estimated time of arrival (ETA) forecast based on current vessel status (location and speed) could result in a balance between speed and delivery of products within the specified time range. A deadline for vessel arrival must be met, failing which the company will be liable to the shipper for a late delivery penalty (Tran and Lam, 2021). However, in order to cut fuel costs, vessels must travel at the slowest feasible speed. Correct ETA prediction can better indicate whether the vessel is on track to arrive at its destination on time or whether a change in speed is required to get it there in the specified time window. If it is estimated that the vessel would arrive early, it can slow its speed to save fuel, reducing its cost while remaining on course.

2.1.2 Terminal and Seaport Operator

Seaport operators or seaport authorities are in charge of the general management of seaports. A seaport is a geographical location where ships dock and unload cargo. Seaports typically have multiple terminals dedicated to specific types of cargo, e.g. (Container, Dry Bulk, Oil). Seaport operators aim to improve the seaport’s total efficiency, including offering separate terminals for different types of cargo, warehousing cargo, and integrating hinterland transit for efficient goods transportation. Uncertainty regarding vessel arrival has a direct impact on Seaport, as the majority of its operations and planning are dependent on vessel arrival (Meijer, 2017). This includes the operation of Terminals. Terminal operations can be classified as landside or seaside. The operation of a seaside facility begins with the mooring of vessels (Wei et al., 2020).

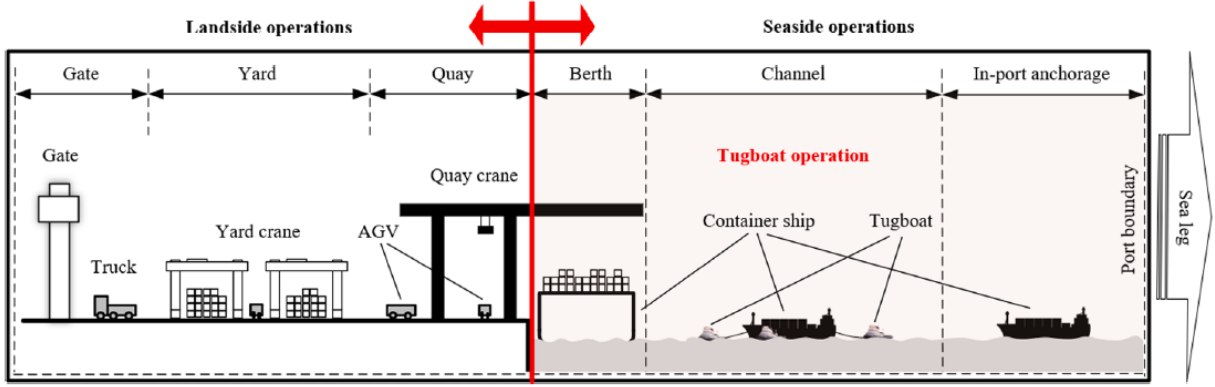


Figure 2.1 landside and Seaside Operations

Source : (Wei et al., 2020)

Mooring of container ships starts with the help of tugboats. Tugboats are required for the following reasons:

- Because the channel and port basin are typically very narrow and congested, ships require the help of tugboats to navigate safely through the port waters (Wei et al., 2020).
- The manoeuvrability of the vessel decreases at the port due to the low sailing speed in the restricted water zones, and the tugboats can assist the ships by pushing or towing them in the correct sailing direction (Wei et al., 2020).

However, the number of tugboats available and the capacity of each tugboat to serve the ships is limited. As a result, the use of tugboats must be adequately managed in order to serve cargo ships successfully and efficiently. Depending on the size of the ship and the cargo, an appropriate berth is assigned to it, where a tugboat assists it in berthing it for unloading with a quay crane. At a port, a container ship may be required to undertake three types of dock operations.:

- Berthing - the process of sailing from the port boundary or anchoring to an assigned berth.

- Shifting - transferring from one berth to another to change the docking position.
- Unberthing - leaving the port or departing from its berth and sailing to open sea (Wei et al., 2020).

As all berthing operations are highly complicated and include a large number of variables, strategic planning is necessary to ensure maximum productivity (Drewry, 2006). To address these challenges created by unexpected vessel arrivals, the port typically allocates more resources than necessary, including people and equipment resources. This leads to increasing expenditures, as labour is the container terminal's largest expense (Fancello et al., 2011). Additionally, the entire planning process is contingent upon the vessel arriving on time according to its estimated arrival time. Any delay in this phase can result in a delay in the entire supply chain process that follows. Therefore, an accurate ETA forecasting tool can be a huge asset to decision-makers.

Following the unloading of the vessels, other activities in the Seaport include transportation to storage, as in a warehouse or yard. Trucks and Automated Guided Vehicles (AGVs) are utilized to transport goods from the query crane to the yard. The transportation of huge amounts of products via these trucks and AGVs needs complicated planning processes (Grunow, Günther and Lehmann, 2007), which might be challenging if the vessel's arrival schedule is unpredictable. Thus, reducing the overall efficiency of Seaport.

2.1.3 Hinterland Transportation

Accurate information about the estimated arrival times of sea vessels is critical for the planning activities of third-party hinterland transportation partners and DP World (the terminal operator of the Port of Southampton), which operates its own hinterland transportation facility. Vessel arrival serves as the beginning point for all hinterland container transport supply chain activity. As with any hinterland shipping, the timely arrival of the vessel is critical. As (Parolas, 2016) stated in his work, delays and projected arrival times

for deep-sea vessels were among the most desired pieces of information for truck operators. Accurate information on vessel arrivals enables hinterland transportation providers to reserve the necessary rail and truck capacity, avoiding over or under-estimation of demand. For example, booking rail transit needs advance reservations, typically several days in advance. If the vessel arrives late and the commodities on board miss their scheduled rail transit, they must be delivered via another mode of transport, such as a truck, which adds to the cost because trucks are less efficient than railroads. As a result, it becomes clear that better containership ETA accuracy can assist reduce the cost of hinterland transport.

2.1.4 Importers

The importer is the one who purchases and receives the cargo contained within the container. The most critical criterion for an importer is timely delivery of their goods at a fair price. As demonstrated in the preceding sections, an ETA prediction tool has the potential to enhance the overall supply chain for container shipping greatly. Because of reduced delays as a result of greater seaport efficiency and cost savings as a result of appropriate resource management.

Efficiency becomes a much more critical factor if goods imported are perishable in nature (Food items and pharmaceutical products). Any delay in regard to perishable goods can have a significant loss for importers as the goods may be spoiled. Also, delays in goods can disrupt the production for importers or make importer hold more goods than needed to circumvent producing issues, which will increase inventory (Xu, Song and Roe, 2011). Thus, a more efficient supply chain process will result in happier importers.

2.2 Current Studies on ETA prediction

In recent years there has been a significant increase in research of ETA prediction for containerships. Predicting the ETA of vessels can be a complex task as there are various

variables associated with it. Thus, many of these researchers employed various machine learning algorithms for the ETA predictor tool (Fancello et al., 2011; Flapper, 2020; Meijer, 2017; Pani et al., 2015; Parolas, 2016) mainly because machine learning models are suitable for handling and understanding vast data that are associated with the voyage of containerships.

Moreover, previous research has demonstrated that machine learning can result in more accurate forecasts for vessels travelling to a port (Fancello et al., 2011; Flapper, 2020; Meijer, 2017; Pani et al., 2015; Parolas, 2016). However, the most similar research done is by (Pani et al., 2015), which only focused on vessel voyages between 2 seaports. Thus, was limiting in nature. However (Parolas, 2016), his research focused on predicting the ETA of vessels using SVM and neural network algorithms that are multiple days away, primarily using AIS (Automatic identification system) data combined with weather data (Meijer, 2017), expanding the study done by (Parolas, 2016) by using K-Nearest Neighbor(KNN) algorithm while incorporating routes data, i.e. route the vessel takes to reach the destination. Both of these researches were focused on Port of Rotterdam.

In the other studies done on ETA prediction, the models that often performed well were: Gradient boosting, Support Vector Machines, and Neural networks, Random Forest (Flapper, 2020; Meijer, 2017; Pani et al., 2015; Parolas, 2016). Moreover, Neural Networks models often performed the best but required much more data compared and computation power compared to other models (Yu et al., 2010).

To summarise, the researches:

Reference	ML Models	Port	Focus
(Pani <i>et al.</i> , 2014)	DT	Cagliari	Short Horizon prediction -24hours
(Pani <i>et al.</i> , 2015)	RF	Cagliari	Only Voyage between 2 port Using Weather data
(Parolas, 2016)	SVM, NN	Rotterdam	Using AIS and Weather Data. Medium Horizon - 7 Days
(Meijer, 2017)	KNN, NN, SVM	Rotterdam	AIS and Route Data
(Flapper, 2020)	Gradient Boosting, SVM	Not Mentioned	General using AIS and Vessel detail

Table 2.1 - Current Studies of ETA prediction: Summarised

2.3 Factors Affecting ETA

Numerous studies demonstrate that weather plays a significant role in maritime delays, as vessel speed is directly related to weather conditions (Lee et al., 2018; Meijer, 2017; Pani et al., 2015; Parolas, 2016). Stormy weather, with its intense winds and waves, can significantly reduce a vessel's speed and fuel economy. Additionally, weather conditions can have an effect on routing considerations, adding to the vessel's delay. While the weather element is evident, there are numerous more hidden factors that might have a significant impact on a vessel's ETA. According to the study (Pani et al., 2015), the second most critical aspect was service, followed by weather. Port rotation, sailing direction, and previous port information are all included in the service. The service indicates the set of ports the vessel loops through and which port it departs from, which is directly related to the distance the vessel must travel to reach its current port. Also, from the works of (Parolas, 2016), it has been established that the distance travelled by vessel has a direct relationship to its ETA. Additionally, as per the findings of (Parolas, 2016), the most significant element affecting the effectiveness of its ETA prediction model was the ETA

information provided by the ship's captain. The ship's captain's estimated time of arrival (ETA) indicates the captain's intentions for how quickly he intends to sail the remaining distance to the port. This indicator cannot be deduced from the other characteristics.

(Parolas, 2016) used AIS data along with Weather data as per his finding, weather data was not much help in his ETA prediction model. Although this appears counterintuitive, he explains that AIS data includes the following: Current speed, Change in speed over the last three hours, and Average speed over the last 12 hours. This data is directly related to the weather conditions under which the vessel is travelling. Furthermore, another study (Pani et al., 2014) demonstrated that the day of the week has an effect on ETA. Ships coming on weekends have a greater chance of being delayed, while those arriving on weekdays have a higher probability of arriving on time.

To summarise, all the factors that affect the ETA of vessels are as follows:


Factors	Reference
ETA provided by ship's Captain 	(Pani <i>et al.</i> , 2015; Parolas, 2016)
Weather Data: Wind Speed, Current Speed, Wave Direction, Wave Height	(Lee <i>et al.</i> , 2018; Meijer, 2017; Pani <i>et al.</i> , 2015; Parolas, 2016)
AIS Data: Distance to Cover, Current Speed, Avg Speed, Current GPS Position	(Meijer, 2017; Parolas, 2016)
Service of the Vessel	(Pani <i>et al.</i> , 2015)
Weekdays	(Pani <i>et al.</i> , 2014)

Table 2.2 - Factor affecting Vessel ETA with Reference

2.4 Research Gap

As a result of the increase in worldwide container traffic, port efficiency has become a hot topic, as demonstrated in the preceding section. Numerous studies and research are being conducted on various seaports, primarily on ETA prediction, as the majority of seaport activities are based on vessel arrival. However, the majority of researchers obtained a somewhat different outcome from the study. Thus, it is self-evident that a single ETA prediction model is inapplicable to all seaports due to the fact that each port has unique characteristics such as geographic location, weather, vessel routes, and operators. Additionally, many other elements that influence ETA over the long term have not been investigated, as the majority of the study has been on short term ETA prediction. One such variable is the vessel turnaround time (TAT) in prior ports; as demonstrated by (Stepec et al., 2020), there is considerable variation in the vessel turnaround time, which has a direct effect on the vessel's departure time from the previous port. As a result, it can have a direct effect on the estimated arrival time at future ports. Moreover, there is no research establishing a relation between vessel ETA and Covid cases.

Concerning the machine-learning model, although a number of popular machine learning methods have been investigated in terms of predicting ETA. However, no extensive research has been done on the performance of ensemble machine learning models. These ensemble models, which incorporate a variety of machine learning techniques, frequently outperform single-method models (Wang, Liang and Delahaye, 2018). Ensemble models frequently combine techniques for learning from historical data and learning about the current situation. Thus, the model can detect trends in historical data and adjust its results to the current situation.

This research differentiates from other research by following ways:

- Concerned about Port of Southampton, as per the knowledge of the writer, there is

no ETA prediction study on ports of UK. As mentioned previously, every port has different characteristics and needs to be studied individually.

- Attempting to forecast the arrival time of the vessel based on the initial port call received by the Port of Southampton. This is the first ETA information the port has received. This is distinct because the majority of the above-mentioned works are focused on the short term, typically five days or fewer. However, in practice, the first port received by a port can be up to 20 days ahead of schedule.
- Attempt to establish relation between Covid cases and vessel ETA with R04 - Is Vessel ETA affected by covid cases?
- Due to the limited availability of data, only vessel characteristics and port call data are used. However, this can show some interesting insights—for example, the ETA prediction model’s accuracy in the absence of AIS and weather data. Additionally, because vessel characteristics and port call data are less complex than AIS and weather data, this approach will require less infrastructure to deploy. Specifically, storage and computation.

Chapter 3 - Methodology

*This **Chapter** details the approach to the project and the methodology used to accomplish the research objective. **Section 3.1** Describes the CRISP-DM process model in general, which will be adopted for this project. The subsequent sections follow phases of the CRISP-DM approach. **Section 3.2** Investigates the understanding and required of this project and links it with Research objectives. **Section 3.3** Explain that data will be a requirement for the project to answer the research question. **Section 3.4** Describes the process of data preparation required for machine learning models. **Section 3.5** Explains the machine learning models selected for this approach and why? **Section 3.6** Describes the performance metric that will be used for model understanding and evaluation.*

3.1 CRISP-DM Process Model

This project adopted the CRISP-DM process model as it provides a standard framework for planning a project. The Cross Industry Standard Process for Data Mining (CRISP-DM) is a six-phase process model that reflects the natural life cycle of data science (Umair Shafique and Haseeb Qaiser, 2014). It is a framework for planning, organising, and executing a data science or machine learning project.

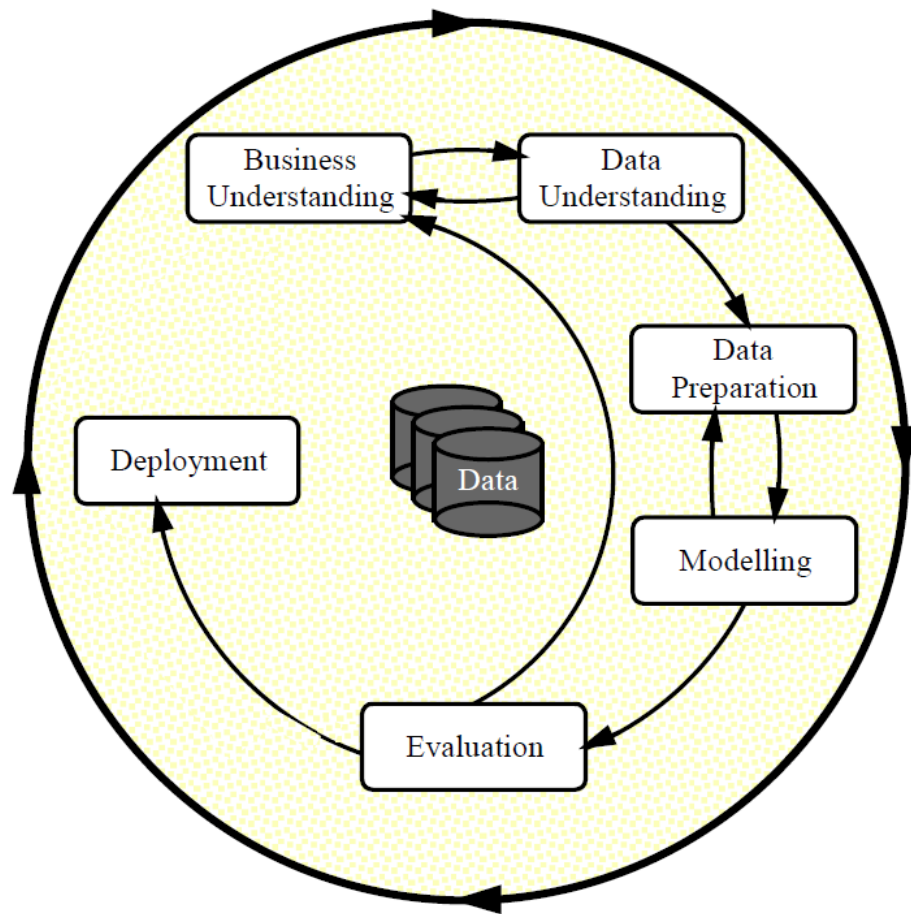


Figure 3.1 - CRISP-DM Process Phase

Source:(Wirth, 2000)

1. Business Understanding - This initial phase focuses on understanding the project's objectives and needs from a business standpoint and then transforming that information into a data mining problem and a preliminary project plan to accomplish the objectives (Wirth, 2000).
2. Data Understanding - The data understanding phase begins with initial data collection. It continues with activities aimed at familiarising individuals with the data, identifying data quality issues, gaining initial insights into the data, or detecting interesting subsets to generate hypotheses about hidden information (Wirth, 2000).

3. Data Preparation-The data preparation phase encompasses all operations necessary to create the final dataset from which the machine learning model will be trained. Tasks associated with data preparation are likely to be repeated several times and in no particular order. Selecting tables, records, and characteristics, cleansing data, creating new attributes, and transforming data for modelling tools are all tasks (Wirth, 2000).
4. Modelling - This phase involves the selection and application of various modelling approaches and the calibration of their parameters to their ideal values. Typically, multiple techniques exist for the same sort of data mining task. Certain approaches necessitate the use of specific data formats. Data Preparation and Modeling are inseparably linked (Wirth, 2000).
5. Evaluation - Created Models are evaluated, and their performance is tested with various evaluation metrics (Wirth, 2000).
6. Deployment - After model creation and testing, It is deployed in the business environment for its value. The final model evaluation is based on the real-world value it provides (Wirth, 2000).

This project will follow CRISP-DM process phases, and the subsequent sections will explain activities for those phases.

3.2 Business Understanding

For undertaking a data science project, the foremost requirement is to understand the "Why" of the Project. That is, "Why is this project needed, and what utility or benefit can this provide?"

As mentioned previously, the importance of an ETA prediction tool and how it can positively affect the container ship industry by increasing the efficiency of stakeholders in-

volved.

As the project's necessity is clear. To accomplish the project's Primary Object that is creating a tool to predict delay in ETA of container Vessel and also to answer R02-What are the primary factors influencing vessel delays? R03-Which machine learning algorithm produces the most accurate results? R04 - Is Vessel ETA affected by covid cases?

Data on ETA must be mined or collected. According to prior research, historical data such as port call and vessel information, as well as weather and AIS data, have shown importance. Hence they are required for the development of a machine learning-based ETA prediction tool.

3.3 Data Understanding

Section 2.3 revealed the relevant data required for undertaking this project, such as ETA of Ship's Captain, Weather data, AIS Data, Service, and Weekday. So, such data regarding the Port of Southampton were collected.

The data that was used for the thesis was obtained from two sources:

- DP World Southampton (<https://www.dpworld.com/southampton/port-info/wheres-my-ship>), This contains information that was received by port from the vessel along with which berth vessel is and will be assigned to. This public information was recorded in excel sheets every five days from 4 July 2020 to 5 Oct 2021.
- Clarksons World Fleet Register (<https://www.clarksons.net/wfr>) contains data regarding vessel charters. Clarksons is the world's leading shipbroking company, i.e. They provide shipping services by linking shipowners with charters. Clarksons being an industry leader, they have a vast availability of data regarding vessels, and it is

publicly available.

These two sources were selected because, from them, 3 out of 5 factors affecting ETA that was discussed in section 2.3 can be derived. These include ETA provided by the ship's Captain, Service of the Vessel, Weekday's data. However, Due to limited data availability, weather and AIS data were omitted from this study.

Raw Data from DP World Southampton consist of 5020 Row and 10 Columns with Features as follows:

Feature Name	Data Type	Description
Vessel Name	Char	Name of the Vessel arriving at Port
Service	Char	Represent the route vessel follows
Phase	Char	Current Status of vessel. i.e Inbound, Departed etc
ETA	DateTime	Estimated arrival time of Vessel
Berth	Char	Berth that Vessel was assigned at the Port
Side to	Char	Side unload/loading takes place, i.e Left or Right of vessel
ETD	DateTime	Estimated time of Departure
Next port of call	Char	Next Destination
Ref	Char	Reference number of Vessel
Recording	DateTime	Date the data was recorded

Table 3.1 - DPWorld Southampton Data

Raw Data from Clarksons World Fleet Register consist of 5496 Row and 15 Columns with Features as follows:

Feature Name	Data Type	Description
Type	Char	Type of the Vessel
Name	Char	Name of the Vessel
Size	Int	Size of the Vessel
Unit	Char	Unit of Size ie TEU
Dwt	Int	Deadweight tonnage, ie How much Ship can carry
GT	Int	Gross tonnage
Flag	Char	Country of registration
Built	Int	Built Year
Month	Int	Built Month
Builder	Char	Builder of the Vessel
Owner Group	Char	Owner of the Vessel
Status	Char	Current Serive Status, i.e In-service, Repair etc
Alternative Fuel Types	Char	Alternate fuel type Vessel Supports
SOx Scrubber Status	Char	Vessel using SOx Scrubbers, Fitted or No
Eco - Electronic Engine	Char	Eco Engine Type Vessel is using

Table 3.2 - Clarksons World Fleet Register Data

Also, UK News cases Covid data from (<https://coronavirus.data.gov.uk/details/download>) a UK government repository for Covid Cases were collected. Which will help to Answer R04 - Is Vessel ETA affected by covid cases? And establish a relation between ETA and new covid cases. UK Covid Data consists of 636 Rows and 3 Columns.

Features as Follows:

Feature Name	Data Type	Description
Date	DateTime	Date of Recording
new_cases	Int	New Cases on that day
total_cases	Int	Total cases till date

Table 3.3 - UK Covid Data

Although from literature review section 2.3, It is evident that some features will have importance, such as Service, ETA and weekday- Which can be derived from ETA, that will be discussed in the 3.4 Data Preparation section. However, this project will try to find the correlation and importance of all of these features listed above.

For initial data understanding expiration, correlation of continuous variables will be identified using Pearson correlation coefficient and, which is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationship or correlation. Further correlation between variables will also be explored with the scatter plot with a simple linear regression line. A Scatter plot is a graphic plot that uses cartesian coordinates to plot two variables from a set of data. Regression line, Simple linear regression is a statistical technique that enables the analysis of the relation associations between two continuous (quantitative) variables.

3.4 Data Preparation

3.4.1 Correcting Data format

Raw data from DP World Southampton had formatting irregularities, which need to be fixed in order to feed into the machine learning models. Thus, Jupyter Notebook with Python interpreter along with Pandas, Numpy packages were to handle and manipulate data.

	A	B	C	D	E	F	G	H	I	J
1	Vessel Name	Service	Phase	ETA	Berth	Side to	ETD	Next port of call	Ref	Recording
2	NYK WREN	FP2	DEPARTED	25-06-20	SCT2	Starboard	28-06-20	Jeddah, Saudi Arabia	G22 /	04-07-20
3				11:03			17:29		H22	04-07-20
4	BG SAPPHIRE	BGLP3N	CLOSED	24-06-20	SCT1	Portside	25-06-20	Cork, Ireland	23L /	04-07-20
5				14:40			2:36		23L	04-07-20

Table 3.4 - Raw format of DP World Southampton Data

Each Entry was occupying 2 rows as Time was separated from Date in ETA and ETD Columns.

	A	B	C	D	E	F	G	H	I	J
1	Vessel Name	Service	Phase	ETA	Berth	Side to	ETD	Next port	Ref	Recording
2	NYK WREN	FP2	DEPARTEC	25-06-20 11:03	SCT2	Starboard	28-06-20 17:29	Jeddah, S	G22/H22	04-07-20
3	MSC ALBANY	EPIC1	INBOUND	06-07-20 19:30	SCT3	Starboard	07-07-20 7:30	Rotterdam	R27/S27	04-07-20
4	QUEBEC EXPRESS	GEX2	INBOUND	07-07-20 7:00	SCT3	Starboard	07-07-20 19:00	Antwerpe	V25/W25	04-07-20
5	BG EMERALD	BGLP3N	INBOUND	07-07-20 7:00	SCT1	Portside	08-07-20 7:00	Cork, Irela	27N/27N	04-07-20
6	YM WELLSRING	FP2	INBOUND	07-07-20 14:00	SCT2	Starboard	09-07-20 2:00	Jeddah, S	G23/H23	04-07-20

Table 3.5 - After fixing the formatting irregularities

3.4.2 Feature Engineering

This project revolves around predicting the Delay of the Vessel. Thus, Delay is calculated by:

$$\text{Delay} = \text{Estimated Time of Arrival (ETA)} - \text{Actual time of Arrival (ATA)}$$

Thus, creating ATA features in the data is a necessity. This process can also be called Feature Engineering. Feature engineering is the process of transforming raw data into features that more accurately represent the underlying problem for predictive models, resulting in increased model accuracy for previously unseen data (Nargesian et al., 2017). In the data, if the Phase is 'INBOUND', then the ETA is the estimated time of arrival. However, if the Phase is "ARRIVED", 'COMPLETE', 'DEPARTED', 'WORKING'" then ETA Column represent the actual arrival time. So, These "'ARRIVED', 'COMPLETE', 'DEPARTED', 'WORKING'" were all renamed to 'ARRIVED' for Simplicity. Phases such as 'CANCELED', are dropped as those meant that the voyage of the vessel did not make it to the Port of Southampton.

	A	B	C	D	E	F	G	H	I	J
1	Vessel Nam	Service	Phase	ETA	Berth	Side to	ETD	Next	Ref	Recording
319	RAGNA	UNIF	INBOUND	17-05-21 7:30	SCT1	Starboard	18-05-21 4:00	Belfast, U20F/20F		15-05-21
336	RAGNA	UNIF	INBOUND	29-05-21 22:00	SCT1	Starboard	30-05-21 18:00	Belfast, U22F/22F		20-05-21
351	RAGNA	UNIF	INBOUND	23-05-21 19:30	SCT4	Starboard	24-05-21 16:30	Belfast, U21F/21F		20-05-21
369	RAGNA	UNIF	DEPARTED	17-05-21 12:20	SCT1	Starboard	18-05-21 4:24	Belfast, U20F/20F		20-05-21
392	RAGNA	UNIF	INBOUND	29-05-21 22:00	SCT1	Starboard	30-05-21 18:00	Belfast, U22F/22F		25-05-21
414	RAGNA	UNIF	DEPARTED	23-05-21 12:45	SCT4	Starboard	24-05-21 10:00	Belfast, U21F/21F		25-05-21
485	RAGNA	UNIF	INBOUND	31-05-21 19:00	SCT1	Starboard	01-06-21 19:00	Belfast, U22F/22F		30-05-21
516	RAGNA	UNIF	INBOUND	05-06-21 7:00	SCT1	Starboard	06-06-21 7:00	Belfast, U23F/23F		30-05-21
534	RAGNA	UNIF	DEPARTED	01-06-21 1:30	SCT1	Starboard	01-06-21 19:45	Belfast, U22F/22F		05-06-21
589	RAGNA	UNIF	INBOUND	06-06-21 7:30	SCT1	Starboard	07-06-21 7:30	Belfast, U23F/23F		05-06-21
593	RAGNA	UNIF	INBOUND	19-06-21 19:30	SCT1	Starboard	20-06-21 19:30	Belfast, U25F/25F		05-06-21
702	RAGNA	UNIF	INBOUND	12-06-21 19:30	SCT1	Starboard	13-06-21 19:30	Belfast, U24F/24F		05-06-21

Table 3.6 - Same Vessel Name multiple phases with different Ref

From Table 3.6 Same Vessel can have different inbound times recorded on the same date, but they have different Ref. From that, It derived that "Vessel Name + Ref" can be used to identify the Vessel uniquely. Thus, another Feature was created Named "Ship ID "to sort and find the ATA. The Vessel's first port call, i.e. first INBOUND phase, was kept, and all other INBOUND phases were dropped, and the same for the 'ARRIVED' phase, only one was kept because it was needed for ATA time.

Also, some vessels had no INBOUND Entry but had DEPARTED and vice versa. They were dropped as it is impossible to derive ATA or ETA for them.

	A	B	C	D	E	F	G	H	I	J
1	Ship_ID	Service	Phase	ETA	Berth	Side to	ETD	Next	Ref	Recording
1644	RAGNA - 20F/20F	UNIF	INBOUND	17-05-21 7:30	SCT1	Starboard	18-05-21 4:00	Belfast, U20F/20F		15-05-21 0:00
1645	RAGNA - 20F/20F	UNIF	ARRIVED	17-05-21 12:20	SCT1	Starboard	18-05-21 4:24	Belfast, U20F/20F		20-05-21 0:00
1646	RAGNA - 21F/21F	UNIF	INBOUND	23-05-21 19:30	SCT4	Starboard	24-05-21 16:30	Belfast, U21F/21F		20-05-21 0:00
1647	RAGNA - 21F/21F	UNIF	ARRIVED	23-05-21 12:45	SCT4	Starboard	24-05-21 10:00	Belfast, U21F/21F		25-05-21 0:00
1648	RAGNA - 22F/22F	UNIF	INBOUND	29-05-21 22:00	SCT1	Starboard	30-05-21 18:00	Belfast, U22F/22F		20-05-21 0:00
1649	RAGNA - 22F/22F	UNIF	ARRIVED	01-06-21 1:30	SCT1	Starboard	01-06-21 19:45	Belfast, U22F/22F		05-06-21 0:00

Table 3.7 - INBOUND and ARRIVED for each unique Vessel

In Table 3.7 Ship ID with INBOUND phase represents the ETA and ARRIVED represents the ATA. A new column was created named ATA then ETA from the row below i.e ARRIVED Phase was added to ATA. Then row with ARRIVED Phase was deleted. From ETA and ATA another feature was created called "Delay" Which stores the difference between ETA and ATA in minutes.

	A	B	C	D	E	F	G
1	Ship_ID	Vessel Name	Service	Phase	ETA	ATA	Delay
823	RAGNA - 20F/20F	RAGNA	UNIF	ARRIVED	17-05-21 7:30	17-05-21 12:20	290
824	RAGNA - 21F/21F	RAGNA	UNIF	ARRIVED	23-05-21 19:30	23-05-21 12:45	-405
825	RAGNA - 22F/22F	RAGNA	UNIF	ARRIVED	29-05-21 22:00	01-06-21 1:30	3090
826	RAGNA - 23F/23F	RAGNA	UNIF	ARRIVED	05-06-21 7:00	07-06-21 4:53	2753
827	RAGNA - 24F/24F	RAGNA	UNIF	ARRIVED	12-06-21 19:30	13-06-21 15:00	1170
828	RAGNA - 25F/25F	RAGNA	UNIF	ARRIVED	19-06-21 19:30	20-06-21 23:45	1695

Table 3.8 - Table with ETA,ATA and Delay

DP world data was cleaned and prepared. Then The Clarkson world fleet data was merged with it. Clarkson World fleet data was already almost clean. However, Many Rows had empty values regarding 'Alternative Fuel Types', 'SOx Scrubber Status', 'Eco - Electronic Engine' columns. They were assigned 0.

In UK Covid data, new cases per week were calculated and merged with DP world. New Cases were taken into account because, as per the initial hypothesis, covid may affect the delay of ETA because the Government tends to restrict activities to stop the covid spread, thus affecting the port's performance. Further explaining will be given once a relationship is established in section 4.4.

The DateTime data type cannot directly feed into the machine learning model (Lee, 2019). Thus, Year, Month, Day, Hour, Weekday must be extracted from ETA. Hence 5 More features were created.

3.4.3 Data Normalization

As different machine learning models will be applied, it is best practice to normalize the continuous data. Normalization is the process of converting the values of numeric columns in a dataset to a similar scale without distorting the ranges of values or losing data. Normalization is also essential for some algorithms to accurately represent the data (Jo, 2019). Min-Max normalization for continuous data in the dataset.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Min-Max normalization formula

3.4.4 Categorizing Target Variable

Delay being a continuous target variable, was categorised into three categories OnTime, Early, Late. "OnTime" refers to vessels that reached the Port at the scheduled time that is at its announced ETA. The "Early" mean vessel arrived before the ETA, and "Late" means it was delayed, i.e. Arrived after the ETA.

However, when categorising the Delay variable, it was categorised based on a range: If the delay was ± 12 hours, then it was categorised as OnTime (24hr range or a Day), If the delay was greater than 12 hours, then it was categorised as Late. If the vessel arrived more than 12 hours early, it was Early.

This was done because:

- As per Sea-Intelligence Global Liner Performance, August 2018 Report average delay of vessels is around 3.5 days. This is a massive uncertainty for the stakeholder (Sea-Intelligence, 2018). However, if our model can predict the accuracy within a day(24hr) as it was done with the Ontime Category, it can prove to be valuable to the stakeholders as it will decrease the uncertainty by 2 days, which is quite substantial.
- Furthermore, due to the limited availability of data and missing AIS and weather data, a classification was used to increase the model accuracy and create a reliable model.

3.4.5 Encoding Data

Scikit-learn python package is used for modelling. Scikit-learn package provides an easy-to-use pre-built machine learning model while providing customizability with user-customizable model parameters. Parameters refer to how the machine learning model will learn or train. However, Scikit-learn requires data to be in Label encoded for the Decision tree and Random forest model. And, One Hot Encoded for SVM and Neural network.

Label Encoding: Encode categorical labels with a value between 0 and $n - 1$, n being the number of categories in that column. This is great if the values have ordinal relation between them. Example Good, Great, Excellent, can be Labeled encoded into 0,1,2. As these numbers perfectly represent the Categorical variable, however, if there is no relation between this variable, then Label Encoding can be misleading to the machine learning model as it may try to make relations such as higher is better or vice versa. This Encoding technique is great for decision trees and Random forests, where the model does not take ordinal value into consideration in its learning process.

One Hot Encoding: This encoding is perfect for categorical variables that have no order or relationship. This is where the categorical variables variable is removed, and a new binary variable is added for each unique categorical value. It creates a new column based on the unique categorical value of a feature. Example:

Pet	Cat	Dog	Turtle	Fish
Cat	1	0	0	0
Dog	0	1	0	0
Turtle	0	0	1	0
Fish	0	0	0	1
Cat	1	0	0	0

Figure 3.2 One-Hot Encoding

Source : (Koelpin, 2020)

Categorical variables must be One-Hot Encoded before feeding into SVM and Neural networks.

Thus, 2 datasets were created, one using Label encoding for the Decision tree and Random Forest. Another is the One-Hot encoding data set for SVM and Neural networks. After the whole data preparation process:

Label encoded data sets consisted of 876 rows and 22 Columns.

One-Hot Encoded Data sets consisted of 876 rows and 503 Columns.

Label encoded data sets:

Feature Name	Data Type	Description
Delay	Char	Category:Early, OnTime, Late
Size	Float	Size of Ship
Days_Before_ETA_Info	Float	Days before ETA info
New_Cases	Float	New Cases in UK at the Week Vessel Arrived
ETA_Year	Int	Year of ETA
ETA_Month	Int	Month of ETA
ETA_Day	Int	Day of ETA
ETA_Weekday	Int	Weekday of ETA
ETA_Hour	Int	Hour of ETA
Built_Year	Int	Built Year of Vessel
Built_Month	Int	Built Month of Vessel
Alternative_Fuel	Int	Alternate fuel type Vessel Supports
SOx_Scrubber	Int	SOx Scrubbers status of Vessel
Eco_Engine	Int	Eco Engine Type
Vessel_Name	Int	Vessel Name
Service	Int	Service
Builder	Int	Who Built the Vessel
Owner	Int	Onwer or Operator of Vessel
Flag	Int	Country of registration
Berth	Int	Berth that Vessel was assinged at the Port
Side	Int	Side unload/loading takes place
Next Port	Int	Next Destination

Table 3.9 - Label Encoded Data set features

All Char features were converted to Int because data was Label Encoding, as it converts categorical value to numerical value as previously explained. Float datatype means features were normalized.

3.5 Modeling

The fundamental purpose of this research is to create machine learning models, which aids in achieving the primary project goal. That is, to develop a model to estimate the arrival time of a containership. The model will be trained on previously cleaned, engineered, and encoded data.

The modelling process will achieve in predicting the Delay variable, which was feature

engineered in the data preparation process. Delay can also be called the Target variable or the dependent variable. Also, the delay was converted into a categorical variable with three classes, Early, OnTime, Late. Thus, a classifying machine learning approach will be adopted. All the features or independent variables will be used to train the model to predict the Delay Class, i.e., Early, OnTime, Late.

Four machine learning models will be developed and compared. These are Decision Tree, Random Forest, Support Vector Machine (SVM), and Neural Network, especially Multi-layer Perceptron (MLP). The subsection of this section will explain each of these models and provide reasons for its selection.

All models will be compared on the basis of their mean accuracy in k fold cross-validation. In K-fold Cross-validation, all of the data is used. Data is divided into k numbers of groups, then take 1 group as test data and the rest of remaining groups as train data. If 10 k fold is selected, then 10 different model training will be performed, hence giving 10 different accuracy results. The model will be performed based on the mean accuracy of 10 k-fold scores. K-Fold allows all data to be used as test and train data. This reduces the model's variability, which is not possible when selecting a traditional train/test split. 10 value of K-fold was selected as it provided most optimal results (Marcot and Hanea, 2020).

3.5.1 Decision Tree

The decision tree is a straightforward classifier among the most often used classification techniques (Steinbach, Tan and Kumar, 2005, p.145). The visual representation is in the shape of a tree, which is why it is called a Decision "Tree." It is a flow chart-like diagram in which each node represents a test on an attribute, each branch of the tree indicates the outcome of the test, and each leaf node provides a label for the class being

tested for. In theory, an exponentially large number of trees with different combinations can be formed using a given collection of features. Some trees are accurate, and others are inaccurate, depending on the approach used. As a result, to create an accurate tree, efficient methods and approaches are utilized (Steinbach, Tan and Kumar, 2005, p.151). Popular methods for creating accurate and optimized trees are Gini Impurity and Entropy. The Gini Impurity or Entropy approach to decision trees is the most often used approach today. Decision trees that are based on entropy nodes are split based on entropy. Entropy is a measure of how random the information is—the greater the entropy, the greater the amount of randomness in the data. The Gini impurity is a statistic that assesses the likelihood that any piece of a dataset will be incorrectly identified when it is randomly classified.

$$Gini(t) = 1 - \sum [p(j|t)]^2$$

(NOTE: $p(j | t)$ is the relative frequency of class j at node t). Using the above formula, Gini of each child node is calculated. When a node t is split into k partitions (children), the quality of split is computed as:

$$GINI_{Split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

where,

- k = number of children nodes
- n_i = number of records at child i
- n = number of records at node t

Decision tree model for this project will select the lowest Gini impurity for splitting of nodes.

Advantage of decision tree:

- Simple, and less resource intensive
- Data normalization is not required
- Less data preprocessing required

Disadvantage of decision tree:

- Overfitting - as only one tree is created, it tries to fit all the data from the data set in a single tree
- Affected by Noise in data,
- Not Suitable for Large data sets - With more data, one single tree will grow more complex to fit all data.

The decision tree was selected for its simplicity and easy understanding and works well with limited amounts of data. Also, Decision tree classifiers from the scikit learn package include a function to view the feature importance, which will help understand features affecting the ETA, Which will help solve the R02-” What are the primary factors influencing vessel delays?”.

Parameters for Decision tree Model for Scikit-learn package:

- `criterion = Gini` :Gini impurity was using for node splitting,
- `Random_state = 0`, Was set to have consistency in each run
- `max_depth = None`, Tree will be grown until all leaves are pure or only 2 samples in leaf node are left.
- All other parameters were left to default values

3.5.2 Random Forest

Random forest (RF) is a classification and regression method that uses ensemble learning to make decisions. When using a random forest model, prediction is accomplished by creating many decision trees, hence the name "forest." Each tree in the random forest spits out a class prediction, and the class that receives the most votes become our model's prediction (Lee, Ullah and Wang, 2020).

Decision trees are differentiated as they are created using a different subset of training data, this process is also known as bootstrapping. Bootstrapping is a statistical resampling approach that includes a random sampling of a dataset followed by the replacement of the samples. Machine learning models are frequently treated to this method of assessing the uncertainty associated with them. It simply samples data many times with replacement from the original training set in order to generate multiple separate training sets from a single training set source. These processes help RF models to be more generalized and avoid overfitting training data. Trees created by random forest follow the same principle of decision tree and use entropy or Gini for node splitting is discussed in the 3.4.1 section. Results of each tree are considered, and the majority of votes decide the final class label when predicting. This process can be termed aggregation. Thus, Random forest combines both bootstrapping and aggregation, this is often referred to as the Bagging technique

(Lee, Ullah and Wang, 2020).

The project's random forest model will use the Gini measure for splitting its node and will create 100 trees, as more trees can cause overfitting and too less can make it underfit. Increasing the number of decision trees over 100 tends to have negligible gain, while computation power required increases exponentially (Oshiro, Perez and Baranauskas, 2012). RF was selected as its generally an improvement over the Decision tree also as it circumvents the overfitting issue with decision trees. Advantage of Random Forest:

- Reduce Overfitting: As it uses the bagging technique for creating trees, which reduces overfitting and helps in making a more general model.
- Data normalization is not required.
- Robust to outliers: Outliers tend to have less impact on the model as many trees are created, and majority voting is used, i.e. aggregation.
- Can perform well with a limited amount of data.

Disadvantage of Random Forest:

- Complexity and longer train time: As many trees are created, making this decking process complex and computation-intensive

Parameters for Random Forest Model for Scikit-learn package: Disadvantage of Random Forest:

- `n_estimators = 100`, 100 decision tree were created
- `criterion: Gini`, Gini impurity was using for node splitting,
- `Random_state = 0`, Was set to have consistency in each run
- `max_depth = None`, Tree will be grown will all leave are pure or only 2 samples in leaf node are left.

- All other parameters were left to default values

And same as decision tree, RF classifier from sci-kit learn packages has function to view feature importance of the trained model.

3.5.3 Support Vector Machine

Support Vector Machine commonly referred to as (SVM), is a supervised machine learning technique that can be applied to classification and regression problems. In the SVM algorithm, each data point from training data is plotted in an n dimension (n is the number of attributes or features in training data). Each value of a feature represents a coordinate in the n dimension space. Classification is performed by finding a hyperplane that separates the two classes based on the coordinate of the features (Ding, Hua and Yu, 2014).

The hyper plane is optimized by maximizing the distance of the data point that differentiate class labels. This distance is referred to as margin. SVM can also deal with nonlinear data by mapping them into higher dimensions (Ding, Hua and Yu, 2014). Thus, creating new vectors, which can be very computationally intensive as the number of features increases. However, most implementation of SVM with packages such as sci-kit learn utilize kernel tricks to efficiently compute calculation. As SVM doesn't exactly need the vectors, only the dot product between the features is sufficient (Scikit-learn, 2018).

Advantage of SVM:

- Efficient at higher dimension, that able to handle multiple features
- Works well with Limited number of data
- Higher performance with a small data set and its more generalized as it tends to avoid overfitting (Ding, Hua and Yu, 2014).

Disadvantage of SVM:

- Perform poorly when data is noisy
- Under performed when the feature is more than a data sample, i.e. rows.
- As SVM work by plotting data points in n-dimension and classifying them by a hyperplane, this leads to no clear explanation of the classification made by SVM.

SVM was selected as it performed well in precious ETA prediction studies, as was mentioned in Section 2.2. Also, this machine learning technique used a different learning algorithm compared to Random first and Decision tree.

Parameters for SVM Model for Scikit-learn package:

- `kernel = rbf`
- `Random_state = 0`, Was set to have consistency in each run
- All other parameters were left to default

3.5.4 Neural Network

A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics how the human brain operates. Like how our brain processes information using neurons, Artificial neural networks use neurons to understand the data fed into them. A simple neural network consists of input layers and output layers. These layers are made up of neurons. The number of neurons in input layers correspond with the number of features or independent variables a data set has, while output layers consist of neurons based on the number classes label of the target variable, or 1 for a regression model. Neural networks with multiple layers are termed multi-layered perceptrons, and the layers between input and output are called hidden layers. Neural networks can either be unsupervised or supervised (Aggarwal, 2018).

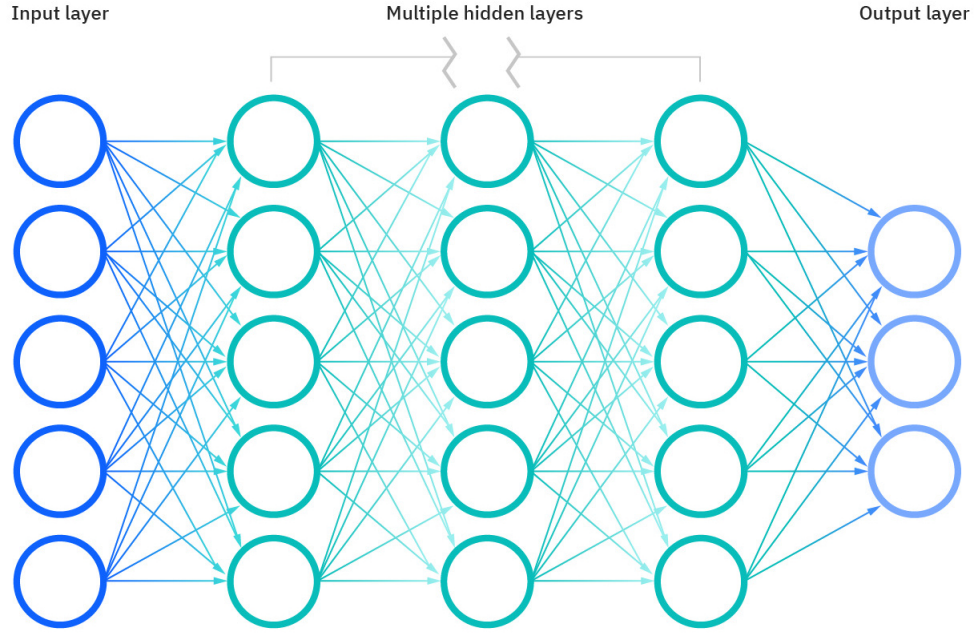


Figure 3.3 - Multilayer Neural Network

Source:(IBM Cloud Education, 2020)

Each dot on the figure represents a neuron and the neuron's activation value is based on the input from the previous layer, for the input layer, it is the feature of the data. Each neuron will have different activation and the weight assigned to them (Aggarwal, 2018). The sum value of each neuron can be calculated as

$$ReLU(w_1a_1 + w_2a_2 + \dots + w_na_n \pm bias).$$

‘a’ and ‘w’ are the activation and weight of the previous layers’ neurons,

‘n’ is the total number of neurons in previous layers,

and bias is used to set the sensitivity of the neuron.

Rectified linear activation function (ReLU) is one of the many functions that affect how neurons activate, other popular functions being Sigmoid and Tanh (Banerjee, Mukherjee and Pasilio, 2019).

The value of weight in each neuron in each hidden layer is adjusted to predict the target variable correctly. A cost function is used to measure the performance of the model. If the

model guesses incorrectly or guesses with low confidence, the cost function is higher. The model tries different values for weight and bias to reduce the cost function. The model uses Gradient descent to converge towards a local minimum by reducing cost function, which is done by adjusting weight and bias appropriately. This process is called backpropagation, as the neural network goes back and changes the parameters (Murugan, 2017).

Advantage of NN:

- Ability to work with incomplete knowledge: after training of model, it can output information even if input have missing information
- Robust to noise in the training data (Aggarwal, 2018).

Disadvantages of NN:

- A large amount of data is required,
- Long train time and computation intense.
- Black Box: No clear explanation on how the output was obtained (Aggarwal, 2018).

From section 2.2 -Current Methods on ETA prediction, it is quite evident that the neural network is reliable in predicting Vessel's ETA. Also, this learning method is also different from previously mentioned machine learning models. Thus, the neural network was selected.

Here a Multilayer perceptron classifier from the sci-kit learn package was used. Multilayer perceptron was created with only 1 hidden layer with 200 neurons. 1 hidden layer was selected based on the research of (Heaton, 2008) because multiple layers show only slight improvement only when using a large set of data, however as our data set is limited with 876 rows, only 1 layer was used. Another critical decision in neural networks is the number of neurons, As high numbers may lead to overfitting of data, whereas low can cause underfitting. Optimal practice suggests the number of neurons should be between

the number of features and the number of target class labels (Heaton, 2008). As we have 503 features (Because of OneHot encoded Data set - Section 3.4.5) and 3 target variables. NN with different neurons was simulated, and it was observed that 300 neurons produced the best balance between runtime and accuracy.

Parameters for MLP for Scikit-learn package:

- `hidden_layer_sizes=(300)`, 1 Hidden layer with 300 neurons.
- `Random_state = 0`, Was set to have consistency in each run
- `activation= 'relu'`
- All other parameters were left to default.

3.6 Evaluation

Created models such as Decision Tree, Random, Forest, Support Vector Machine will be analyzed and compared. After, the best model is selected from the mean accuracy score from k fold cross-validation. Then multiple different classification evaluation metrics will be adopted to measure the performance of this model. The best performing model will be selected then and trained, and tested on a 70/30 split of data. From that, more in-depth analysis will be carried using various evaluating metrics such as:

Accuracy: How often classifier predict correctly.

$$Accuracy = (Total\ Correct\ Predicted \div Total\ Prediction)$$

Confusion Matrix is a table for summarizing the performance of a classification algorithm (Xu, Zhang and Miao, 2020). True positive, negative and False positive, negative can be derived from it as model predict 3 class labels, Early, Ontime, Late. Therefore,

the Multilabel Classification metric will be used to understand the performance better. However, because of Multilabel Classification, there are no clear Positive and Negative classes here. Therefore True Positive, True Negative, False Positive, False Negative will be used to represent the performance of each individual class.

True Positive Indicate, How often it correctly predicted Yes, in our case, How often it model Predicted OnTime and Its actual OnTime.

True Negative: How often Models predict No and it was Actually No.

False Positive: How often model wrongly predict Yes, where actually it was No,

False Negative: How often model predicted No, But it was Yes,

From these, the Recall and Precision of each different class will be calculated.

Precision: When model predict yes, how often its correct.

$$Precision = TP \div (TP + FP)$$

Recall: Revels how often actually yes is correctly classified.

$$Recall = TP \div (TP + FN)$$

F1 Score: It is calculated from Precision and Recall. F1 score is the harmonic mean of

precision and recall.

$$F1\ Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

F1 Score is more sensitive to low values of 2 inputs, i.e. Recall and Precision, as a low Score in one of them can result in a low F1 Score (Visa and Salembier, 2014). The above metrics will help in understanding the model's accuracy in predicting each class.

Furthermore, as our models predict Multi-Label classification, we will move to Calculate **Macro** Recall, Precision and F1 scores, which will take the mean of 3 individual classes. To get a better understanding of the model as a whole and to represent class imbalance.

$$Macro\ Precision = \frac{(Precision\ of\ Class\ 1 + Class\ 2 + \dots\ Class\ n)}{n}$$

$$Macro\ Recall = \frac{(Recall\ of\ Class\ 1 + Class\ 2 + \dots\ Class\ n)}{n}$$

$$Macro\ F1\ Score = \frac{(F1\ Score\ of\ Class\ 1 + Class\ 2 + \dots\ Class\ n)}{n}$$

Chapter 4 - Analysis and Findings

*This chapter discusses the feature importance and compares models in order to answer **R02** and **R03**. **Section 4.1** explores data and reveals key findings. **Section 4.2** considers all features, builds and compares 4 ML models, and extracts Feature importance from the best performing model. In **Section 4.3**, feature selection is used to drop misleading and irrelevant features. **Section 4.4** does an in-depth performance analysis of each target class. **Section 4.5** aim to improve model performance further.*

4.1 Initial Data Exploration

Pearson coefficient matrix will be used to identify the relation between continuous independent variables and target variables as discussed in Section 3.3 Data understanding. For the coefficient matrix, the Delay variable was not transformed to Categorical values, i.e. Early, Late, Ontime.

	Delay	Size	Dwt	Size
Delay	1.000000	0.092989	0.077017	0.092989
Size	0.092989	1.000000	0.992783	1.000000
Dwt	0.077017	0.992783	1.000000	0.992783
Size	0.092989	1.000000	0.992783	1.000000

Figure 4.1 - Correlation Matrix of Size and other volumetric metrics with Delay variable

From Figure 4.1, features such as GT, Dwt and Size are highly correlated, which makes sense as larger sized ships will have more Deadweight tonnage capacity and Gross Tonnage. This can be termed multicollinearity. It's been observed that multicollinearity can be an issue while training a machine learning model (Daoud, 2017). Also, out of these 3, Size has the most correlation with Delay. Hence, only Size will be used for modelling purposes. Dwt and GT will be dropped to avoid multicollinearity issues.

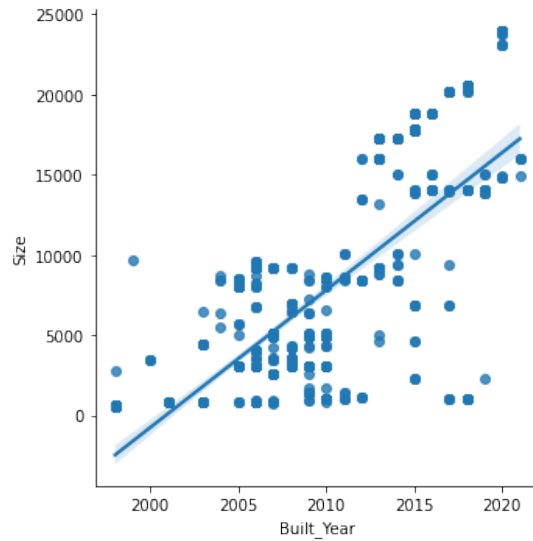


Figure 4.2 - Regression Line, x = Built Year of Vessel, y = Size of Vessel in TEU

Fig 4.2 shows there is a strong positive correlation between the Size and Built Year of the vessel because of the slope of the regression line. This can be explained as recently manufactured vessels are getting larger and larger to increase the efficiency by carrying more goods at a time.

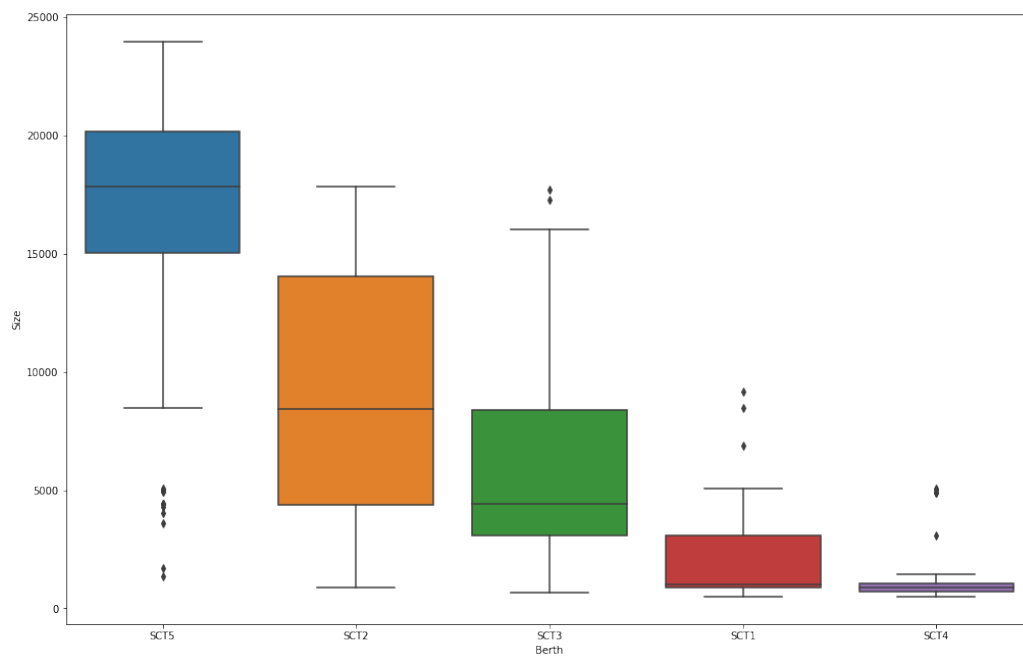


Figure 4.3 Boxplot, x = Berth, y = Size of Vessel in TEU

Also, a boxplot was used to establish a relation between a categorical value (Berth) with continuous value (Size). Figure 4.3 reveals that Berth is assigned to the vessel depending on the size of the vessel. In comparison, SCT4 Berth handles smaller vessels and SCT5 handles bigger vessels. Also, as Berth is categorical, berth numbers are not linear. In boxplot, it was ordered this way for understanding the relationship clearly.

	Delay	Built_Year	Days_Before_ETA_Info	ETA_Year	ETA_Month	ETA_Weekday	ETA_Hour	new_cases
Delay	1.000000	0.102217	0.175336	0.112613	-0.114848	0.139480	-0.005817	0.128013
Built_Year	0.102217	1.000000	0.102848	-0.105188	0.008451	0.049824	0.064705	-0.028814
Days_Before_ETA_Info	0.175336	0.102848	1.000000	-0.086831	0.208492	0.003215	0.066762	0.187386
ETA_Year	0.112613	-0.105188	-0.086831	1.000000	-0.690520	0.071544	-0.002019	0.300959
ETA_Month	-0.114848	0.008451	0.208492	-0.690520	1.000000	-0.027204	-0.041271	0.044926
ETA_Weekday	0.139480	0.049824	0.003215	0.071544	-0.027204	1.000000	-0.024073	0.106013
ETA_Hour	-0.005817	0.064705	0.066762	-0.002019	-0.041271	-0.024073	1.000000	-0.046357
new_cases	0.128013	-0.028814	0.187386	0.300959	0.044926	0.106013	-0.046357	1.000000

Figure 4.4 - Correlation Matrix - Correlation between Delay and Continuous variable

Figure 4.4, States the Pearson coefficient of these continuous variables. Days Before ETA Info has the highest correlation with the delay variable. This indicated that delay is proportionally related to how early ETA information is received. Also, ETA year, Month, Weekday have a decent correlation. That relation shows all aspects of ETA have a relation with delay, which correlates to what was also identified from the Literature Review Section 2.3.

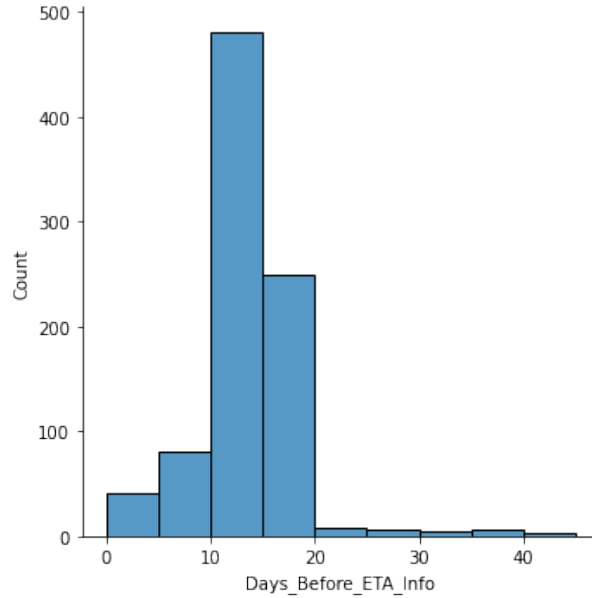


Figure 4.5 - Distribution plot of Days Before ETA Info

Figure 4.5 shows the spread of Days Before ETA info, Majority of port call data were received between 10 and 20 days in advance. As discussed in Data preparation section 3.4, only the first vessel port call was kept. Therefore figure 4.5 indicated vessel's coming to Southampton port usually announce their ETA to port 10 to 20 days in advance.

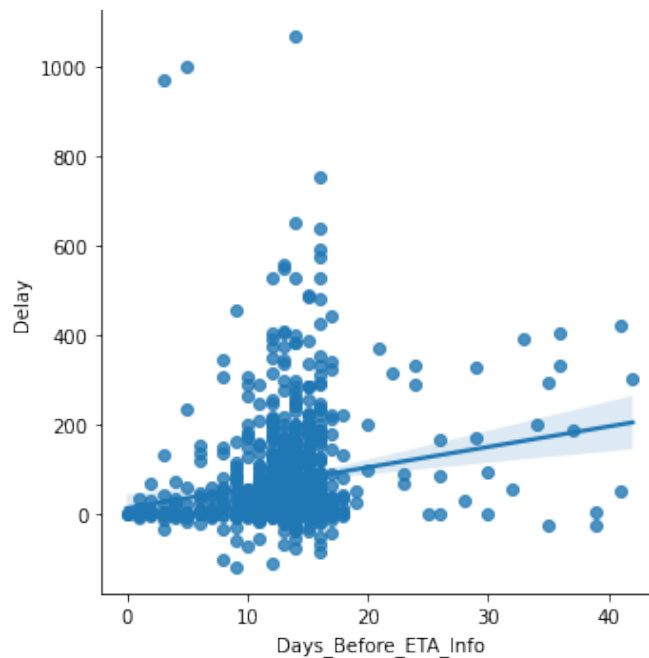


Figure 4.6 - Regression line of Delay and Days Before ETA info

From Figure 4.6, one can see a clear relation on how early ETA information is received and the delay associated with it based on the slope of the regression line. Also, the distribution of data points is concentrated between 10 to 20 days.

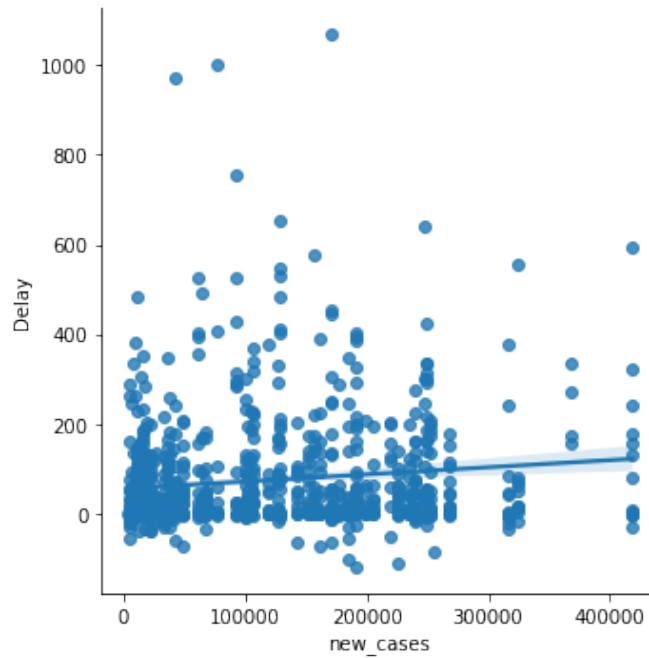


Figure 4.7 - Regression line of Delay and weekly new Covid cases UK

Relation between the number of new covid cases in the UK and Delay was not substantial, as is evident from the slope of the regression line from Fig 4.6. Although as per the slope of the regression line, there is a certain amount of positive relationships. i.e. More new cases can lead to more delay. One assumption can be as new covid cases rise in the UK, nearby country's ports also get affected, thus causing delay because of delay incurred in previous ports as the Government tends to restrict activities when covid cases are rising to stop its spread. However, further study and research are needed to be confident on this hypothesis. Due to the nature of this project and limited data, it's out of scope.

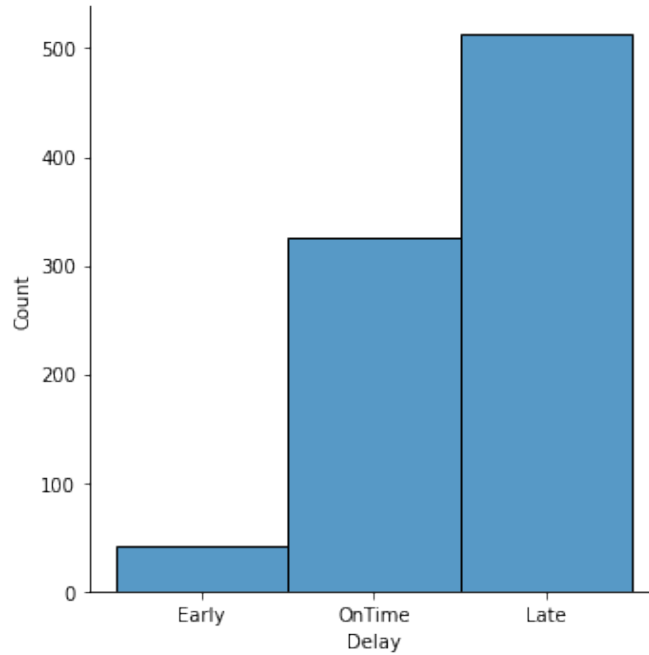


Figure 4.8 - Class labels distribution

Class labels in our data set are unbalanced, especially for Early class labels. The effect of this will be discussed in section 4.4.

4.2 Model with All Features

All 4 models, i.e. (DT, RF, SVM, NN) with all 22 features described in Table 3.9 in Section 3.4.5, were used to calculate their accuracy and f1 by 10-fold cross-validation.

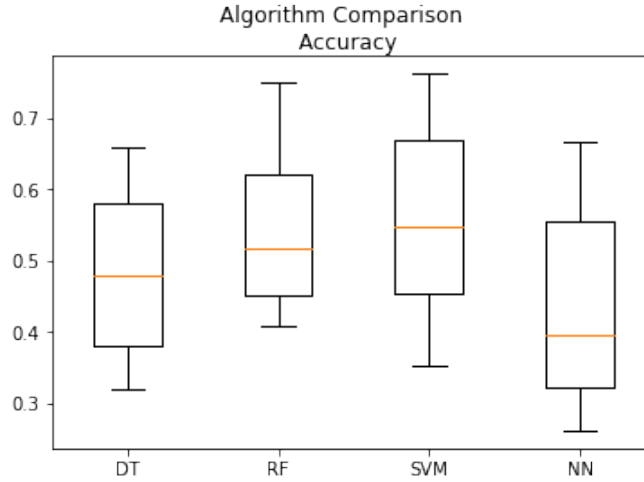


Figure 4.9 - Accuracy of Models - All Features - Boxplot

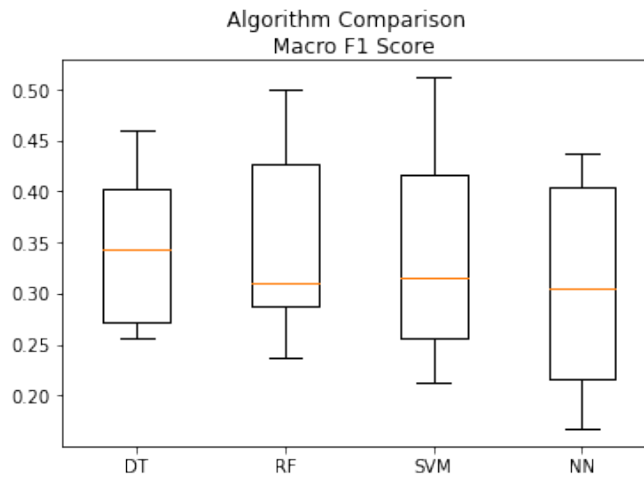


Figure 4.10 - Macro F1 of Models - All Features - Boxplot

Metrics	DT	RF	SVM	NN
Mean Accuracy	0.48	0.55	0.56	0.44
Mean F1 Score	0.34	0.35	0.34	0.31

Table 4.1 - Mean accuracy and F1 score of Model

With all the features, models performed poorly. The decision tree and Neural network are below 50% accuracy, meaning it's predicted wrong most of the time. High differences between accuracy and F1 can also indicate one or more of the class labels is performing far worse in prediction. This will be further investigated in section 4.4. Moreover, Boxplot shows that 10 train/test split that was done by 10-fold cross-validation has a large variance for all the models. Hence Macro F1 Score suffered as few k-fold ran scored low f1 score,

which can be observed from boxplot.

Although all models performed poorly, random forest and SVM were best performing comparatively. Therefore, As the Random Forest scikit-learn model has a function to extract feature importance thus, the RF model will be used to extract the feature importance.

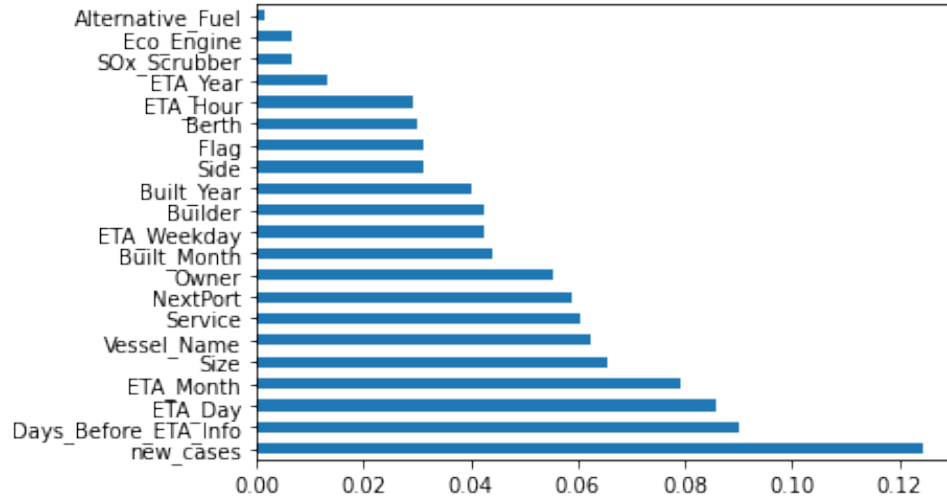


Figure 4.11 - RF Models Feature Importance - All features

From Figure 4.11, it is quite evident that some features have less influence in delays, such as Alternative Fuel, SOx scrubber, Eco Engine, Size, Berth, Flag. Hence, with this knowledge, along with the information from the literature review and correlation identified in section 4.1, we can start feature selection processes. It is a process to reduce the number of input features for machine learning models by eliminating less important features. This can make the model less computation-intensive and usually more accurate.

4.3 Feature Selection

The features were selected based on the insight from literature review section 2.3, Initial data exploration Section 4.1 and feature importance Extracted from Figure 4.11.

4.3.1 Result After Feature Selection

Model performed most well with 'Size', 'Days_Before_ETA_Info', 'new_cases', 'ETA_Month', 'ETA_Weekday', 'ETA_Day', 'ETA_Hour', 'Owner', 'NextPort' Features.

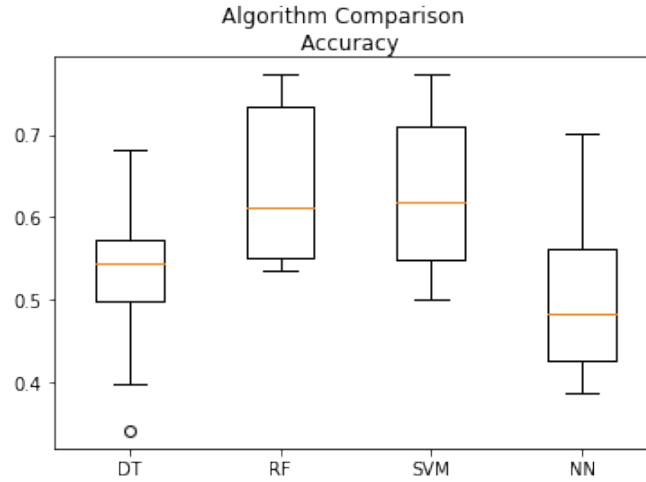


Figure 4.12 - Accuracy of Models - Selected Features - Boxplot

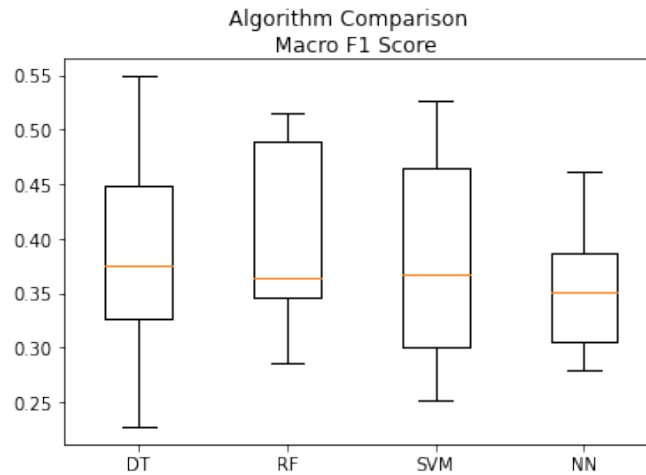


Figure 4.13 - Macro F1 of Models - Selected - Boxplot

Metrics	DT		RF		SVM		NN	
	Before	After	Before	After	Before	After	Before	After
Mean Accuracy	0.48	0.53	0.55	0.64	0.56	0.63	0.44	0.50
Mean F1 Score	0.34	0.38	0.35	0.40	0.34	0.38	0.31	0.36

Table 4.2 - Mean accuracy and F1 score of Models

Random forest, SVM and NN show significant improvement. However, NN and DT are still poor performers. This is due to NN requiring large training data to perform well.

Thus, 876 data was insufficient, and DT usually suffers from overfitting, resulting in poor test accuracy. However, RF, SVM is 64% and 63% accurate in predicting the Delay class. Nevertheless, the F1 score still suffers. Moreover, if you look at the Boxplot spread from figure 4.12,, there is a lot of variance in each 10 k fold variation. Hence, it was a good practice to perform k fold cross validation when comparing different Machine learning models. If the traditional single Train/Split method was used, the result wouldn't represent whole train data. A lot of bias can be present depending on the train/test split (i.e. Which rows were selected randomly for train and test split).

4.3.2 Feature Importance

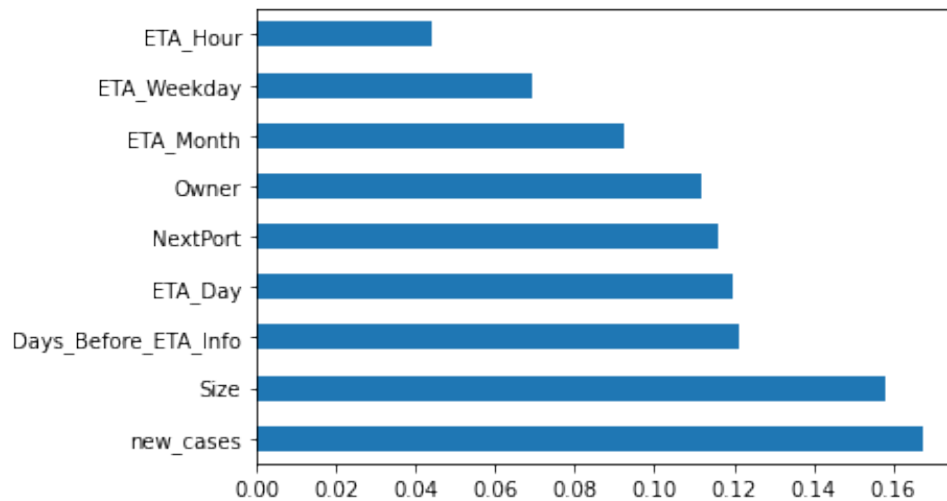


Figure 4.14 - RF Models Feature Importance - Selected Features

Again, we look at the feature importance as per the Random forest model with optimized selected features. ETA__day seems to be most important, combining it with importance with other ETA elements such as Month day, weekday, hour. Therefore, Vessel captain ETA is still the most important factor in predicting Delay.

Size is also important as per this RF model when predicting Delay. It seems larger ships tend to have more Delay. This may be because larger ships have a longer turnaround time

in the last port as larger vessels have more loading and unloading time. This can cause Delay leaving the previous port, thus, making it delayed in subsequent ports.

Days Before ETA Info also is important as earlier the ETA information received the more unreliable the ETA Which leads to more Delay.

New weekly Covid Cases in the UK were also an important factor as they related to new cases nearby, where most ships are coming from. One reason can be that more cases in the previous port can cause vessel delay, although further research and investigation will be needed to draw a valid reason. However, due to limited data availability, it's out of the scope of the project.

The owner, i.e. can be the Operator of the vessel, is also an important factor as per the RF model. This can be intercepted as some operators of vessels tend to be more reliable than others.

Moreover, this finding answers the **R04 - Is Vessel ETA affected by covid cases?** With Yes, there is a relation between the rise in monthly covid cases and Vessel delay.

NextPort may seem an interesting factor. But the next port is closely related to Service. Because Service is Route or set of port vessel voyages. However, the sequence is not always the same with Service. Because of limited data availability, the previous port data is missing. The previous port essentially derives the distance vessels need to travel. The longer the distance, the more probability of Delay, as was discussed in the literature review section. 2.3. Therefore NextPort, with Current Port that is Southampton, may reveal the Previous port most of the time as the 3 port sequence is more reliable than the sequence of the whole Service.

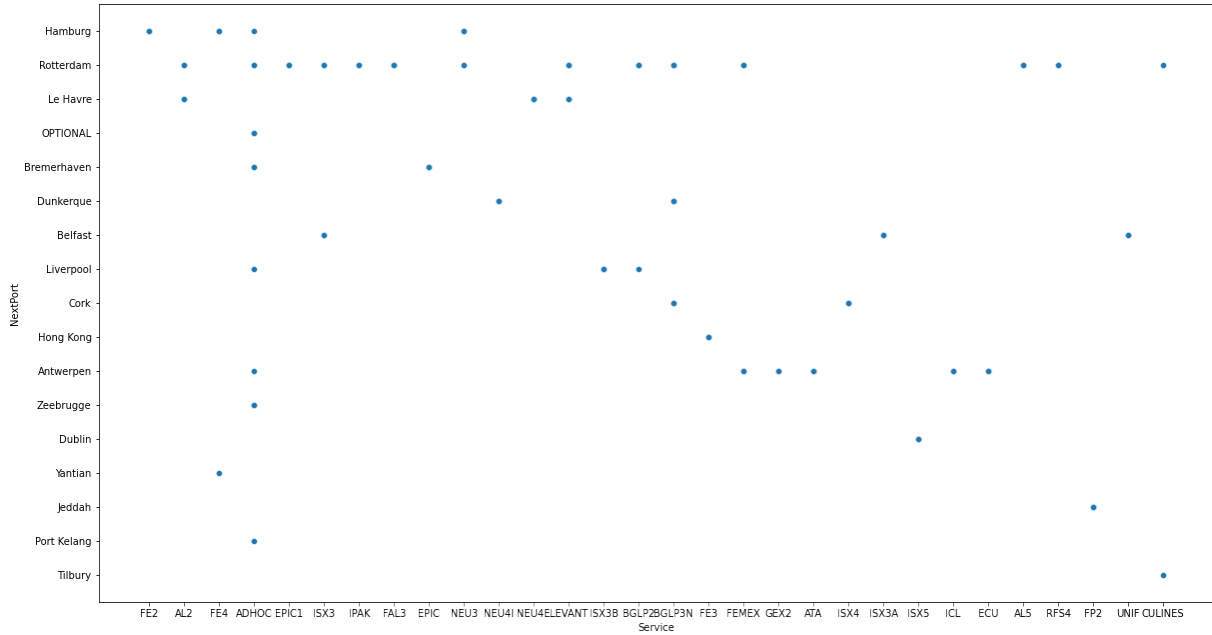


Figure 4.15 - NextPort and Service - Scatter Plot

Figure 4.15 Scatter shows, the same service can have a different next port. Therefore, with the Next port, One can assess the probability of the previous port being the same if 2 of the 3 ports in sequence are the same, i.e. if Southampton and Next Port are the same, the chances of the previous port being constant is high. Therefore, all models perform better with NextPort Compared to Service features.

	Including Next Port		Including Service		Including Both	
Metrics	RF	SVM	RF	SVM	RF	SVM
Mean Accuracy	0.64	0.63	0.62	0.61	0.63	0.62
Mean F1 Score	0.40	0.38	0.38	0.36	0.40	0.38

Table 4.3 - Model Comparison with Nextport and Service Feature

Including Next Port as feature always perform slightly better than Service or Including Both. One of the reasons for Only Next port is better than both is due to Service being misleading as it does not always depict the previous port. Essentially the factor that influences the delay is the distance from the previous port, as previous port data was not

available, so the NextPort feature was used.

4.3.3 Discussion

Random Forest always performed slightly better than SVM and far better than DT and NN. Therefore, from the test result, it is evident that RF is the best model for these available data sets of Southampton port as it provides better accuracy and F1 score. With this result, it can give insight in answering **R02** -" What are the primary factors influencing vessel delays?" as well as answers **R03**-" Which machine learning algorithm produces the most accurate results?" Also with extracting of feature importance relation on New monthly Covid cases and Delay can Be establish which answers **R04** - Is Vessel ETA affected by covid cases?

Answer for R02: Factors affecting Delay for a vessel are:

- Vessel's Captain's ETA
- Weekly New Covid Cases in UK
- Size of the Vessel
- Days Before ETA Information received
- Next Port(As it partially indicates previous port, which was discussed above)
- Owner i.e. Who Operates the Vessels,

Answer for R03: As RF is the best machine learning model for available data and Southampton port because it provided the same accuracy but a slightly higher f1 score than SVM and performed a lot better than DT and NN. Answer for R04: Yes, there is a relation between new monthly covid cases and vessel delay. One hypothesis was highlighted in Section 4.3.2. However, further research needs to be conducted to verify it.

4.4 In depth Class Label based Analysis of RF Model

A confusion matrix will be used to analyze each class label, however for that result from 10-fold cross-validation is not suitable for creating a confusion matrix as 10 different confusion matrices will be created and analysing all of them is not optimal, therefore as we already know RF is best performing model. So, a 70/30 Train/Test split will be used and a confusion matrix will be built up the result of test data. Moreover, this testing will only contain feature that were selected in Section 4.3 Feature Selection 70/30 Train/Test splits data into 616 rows for Training, 265 rows for Testing.

Metrics	Score
Accuracy	1
F1 Score	1

Table 4.4 - Training Accuracy and F1 score of RF Model

Training accuracy for the RF model is 100%, Meaning there was no miss classification regarding training data. As there was no max depth set for tree in random forest, tree grew till it classified all training records hence it is 100% accurate in training data. Now, we investigate test result.

Result of Test Data:

		Predicted		
		Early	Late	Ontime
Actual	Early	0	9	2
	Late	1	133	26
	Ontime	0	43	49

Table 4.5 - Confusion Matrix for Multi-Class Classification

Initial analysis of Confusion matrix, suggest that model is good at Predicting Late Class label, Decent at predicting Ontime class label, and very poor at predicting Early Class

label. This can easily be explained if we look at the distribution of class labels in our data set, Figure 4.8. Out of 876 records, only 40 belong to the Early class. This means the model got very little data to train for the Early Class label. Thus, making it poor at predicting Early Class labels.

	Early	Late	OnTime
True Positive	0	133	49
True Negative	251	51	143
False Positive	11	52	28
False Negative	1	27	43
Precision	0.00	0.72	0.64
Recall	0.00	0.83	0.53
F1-score	0.00	0.77	0.58
Accuracy		0.69	
Macro F1 Score		0.45	

Table 4.6 - Class based performance metrics of Test Data

With other 30% Test data, accuracy seems to improve because random data that was selected may have more bias to them. In comparison, 10-fold cross-validation performs the train test split 10 times and takes the mean. As Early class label data is scarce, it is good practice to remove them and retain the model without an Early class label.

4.5 RF Model with 2 Class Labels

The model was retained with only 2 target Class labels, i.e. OnTime and Late. Early was dropped, thus leaving with 836 Row. The same 70/30 train/split was used. So, 70/30 Train/Test splits data into 585 rows for Training, 251 rows for Testing.

Results for Training Record:

Metrics	Score
Accuracy	1
F1 Score	1

Table 4.7 - Training Accuracy and F1 score of RF Model with 2 Class Labels

Same as before, the training record is 100% correct and perfect f1 score because of max depth for tree in RF model was set to unlimited or there are only 2 samples in leaf node. So it fitted all training record perfectly.

Result for Test Records:

		Predicted	
		Late	OnTime
Actual	Late	133	19
	OnTime	31	68
Performance Metrics			
Recall		0.88	0.69
Precision		0.81	0.78
F1-score		0.84	0.73
Macro F1-score		0.79	
Accuracy		0.8	

Table 4.8 - 2 Classes performance metrics on Test Data

With Only 2 classes, our model started performing a lot better. Specifically, the Recall score of 'OnTime'. It indicated if the Vessel is 'OnTime', and our model also predicted 'OnTime'. This will help the Stakeholders, such as Terminal Operators, be ready and confident with a 69% recall score that This predicted OnTime vessel is going to arrive at OnTime and plan their activities relating to that.

Also, Late has a Recall score of 88%, implying the model predicts with 88% probability that Late predicted Vessel would be late, meaning Terminal Operators can low prioritize

activities related to that Vessel as its more likely to be late. Practical benefits for stakeholders will be further discussed in Section 5.1.

Moreover, overall model accuracy jumped to 80% from 65% with 2 class labels. That means the model can predict 80% of the time the estimated arrival of a vessel within ± 12 hr from the first port call received by Port. which is usually 10 to 20 days in advance. Also, the class-based f1 score and macro f1 score are good, which indicate none of the classes is underperforming, as was the case with the 'Early Class label. However, the model is still better at predicting Late class labels, mainly due to the distribution of train data (Fig 4.16). The model was trained on a more Late class label.

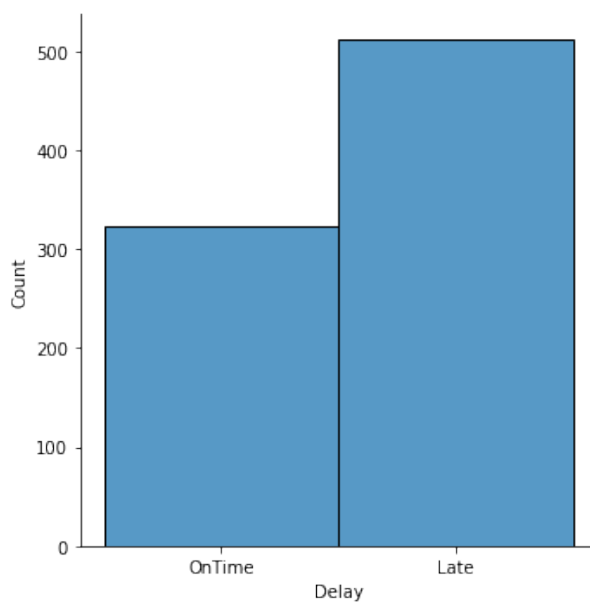


Figure 4.16 - Two Classes label Distribution

Feature importance had some change after using 2 class labels in RF:

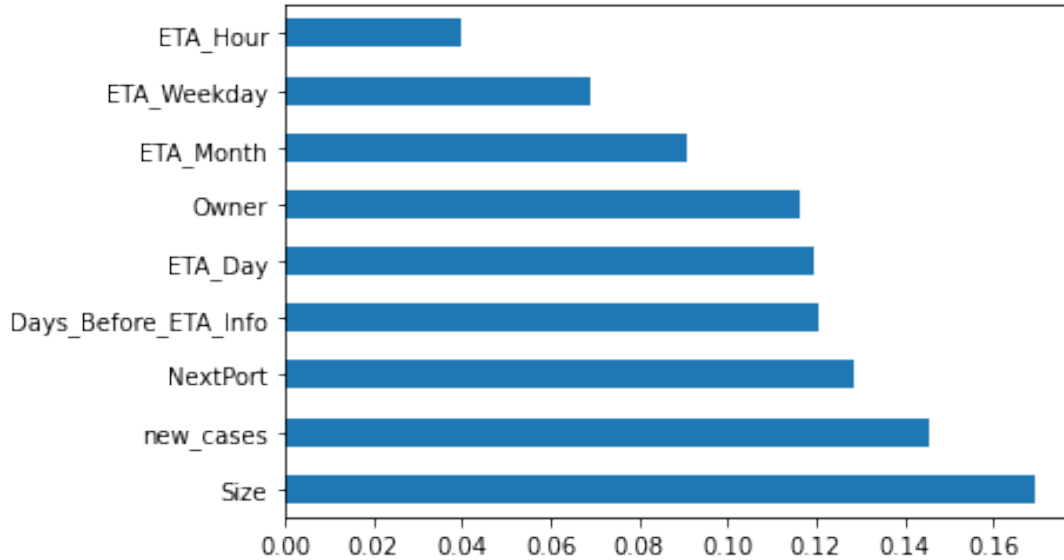


Figure 4.17 - Feature importance of 2 Class Label RF model

As per the RF model, new cases have the highest Gini importance. However, as ETA is divided into various features, its combined importance may be the highest. NextPort and Days Before ETA info still prove to be relevant along with vessel owner. Therefore This model further establishes the relation of ETA, Days before ETA info, NextPort (i.e. Distance from the previous port). While also put forwards new features which affect the Delay of the vessel, such as Size and Owner. and covid cases.

Furthermore, to find the optimal tree number for the RF model, Repeated simulations with different numbers of trees were done (1 to 500 Trees). The same parameters were used except the number of trees created ie `n_estimators`, was changed.

Result of 30% test data as follows:

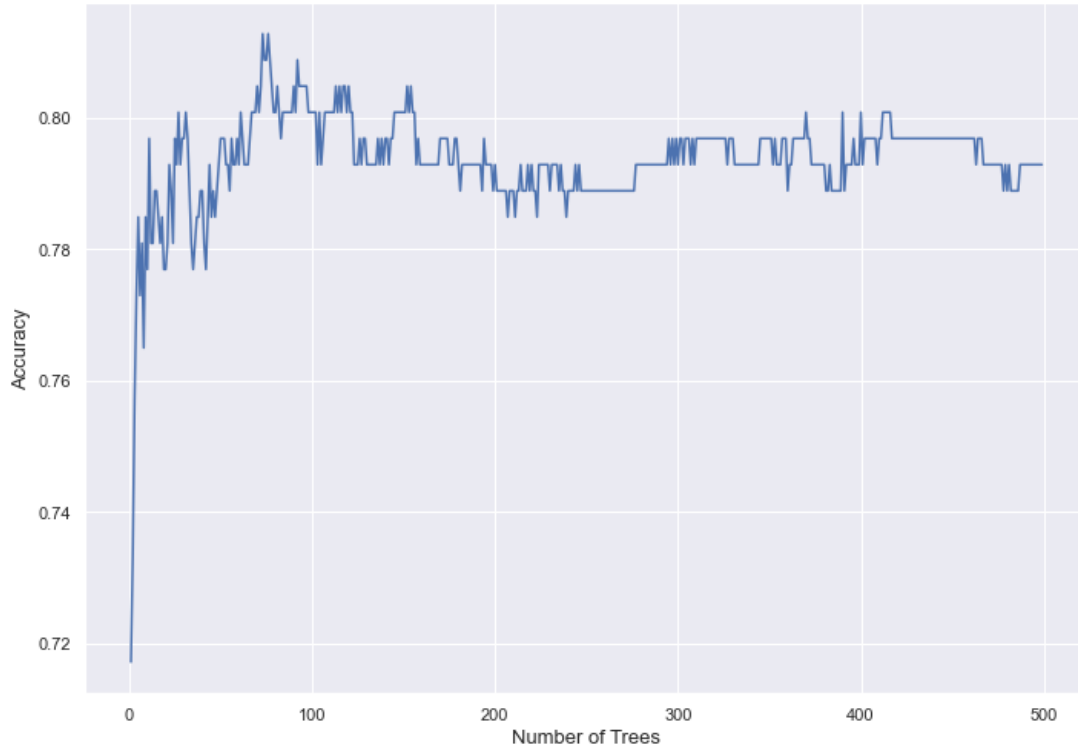


Figure 4.18 - Test Accuracy with Number of trees in RF model

As per simulation, their accuracy tops at around 100 trees, remaining constant with only 1% variance. Even with 500 trees, there is not any noticeable difference. Therefore, initially selecting 100 trees as per the research of (Oshiro, Perez and Baranauskas, 2012) in Section 3.5.1 was optimal.

Chapter 5 - Practical Implementation

*This Chapter discusses the practical implications of the model. **Section 5.1** highlights the effect this model and findings will have on Stakeholders of Container transportation. **Section 5.2** discusses how to implement it on the Port of Southampton and utilized by other Stakeholders.*

5.1 Effect on Stakeholder

The Random Forest model can provide value to stakeholders as many of the stakeholder activities depend on the reliability of the vessel ETA, which was extensively discussed in Section 2.1.

Vessel Operators with this tool can identify if their vessel is on track to reach the scheduled port. If the model is predicting Late arrival to the port, then they can use this information to speed up the vessel to reach on time and avoid late penalty and bad reputation due to being Late. If predicting OnTime, it should continue its course as it is. Moreover, vessel owners can monitor the reliability of other vessel competitors and price their service accordingly.

Terminal Operators, i.e. DP World, will gain the most from this tool, As almost all of the terminal activities are dependent on vessel arrival. These activities include Scheduling tugboats, assigning berths, crane scheduling and workforce management, which was discussed in the 2.1 section. If the model is predicting the vessel is on OnTime, then the Terminal operator can prioritize and plan its activities for that vessel and work on the assumption that this vessel will arrive on the announced ETA. However, if the model is predicting Late, then the terminal operator will make activities related to that vessel a low priority and focus on activities of vessels that are On time. For example, if multi-

ple Vessels are expected to arrive on a particular day, the Terminal operator can look at the model prediction and check how many of these vessels are on time and late. With this information, decision-makers at the terminal can plan ahead of time and distribute resources such as the workforce accordingly. Also, This model can predict the OnTime and Late status of vessels more than 10 days early. This means the Terminal operator can start planning days ahead and implement change if predicting the status of that vessel change with the next port call because the model predicts a new status (OnTime or Late) of each different port call of a particular vessel.

Moreover, Terminal operators can communicate with vessels scheduled to arrive late as per the prediction model and get further detail directly from the vessel operator or the Ship's captain regarding the ETA or delay.

Hinterland transport parties are also dependent on the reliability of vessel ETA as they book transport capacity in advance depending on the TEU arriving at the terminal on that day. Therefore, With OnTime and Late predictions from our model, They will only book the transport capacity of OnTime vessels, and not for Late. This can save a lot of overbooking hinterland transport costs that occur when vessels are late. A smooth and transparent flow of information between Terminal Operator and Hinterland transport parties can make efficient management of activities such as truck fleet routing and schedule.

Seaport efficiency will also improve with increased efficiency of Terminal operators and Hinterland transports as more efficient trade will happen across the supply chain.

Importers can monitor if their cargo is arriving on time or late, which will help them to plan and manage their activities related to that arriving cargo (such as production, storage etc.), moreover as this tool in increasing the efficiency and reducing the cost of other stakeholders in the supply chain process, which may lead to lower shipping prices

for the importer.

5.2 Implementing at Port of Southampton

As this model require only 6 Features to predict

- Days Before ETA Information received,
- New Covid Cases in UK,
- Vessel's Captain's ETA,
- Owner, i.e. Who Operates the Vessels,
- Next Port
- Size of the Vessel,

However, New Covid Cases in the UK variable is redundant because people around the world are getting vaccinated (Covid-19), and the world is preparing for the post-covid era. Therefore, New cases variables for this model will be redundant in predicting future delays, as this feature won't have any effect. Therefore, the model should be trained on new post covid data without New cases features. However, if in future there is another pandemic. This project can provide a relation between new pandemic cases and vessel arrival delay.

That leaves with only 5 features, and out of 5, 2 features (ETA from port call and NextPort) are already being collected and stored by DP World (Terminal Operator of Port of Southampton) and Days Before ETA Information received can be easily derived from it as it was done in Data Preparation Section 3.4. Furthermore, for Vessel size and Owner, this can be easily obtained from Clarksons World Fleet Register, which is publicly

available.

Therefore, the Port of Southampton should undertake real-time testing of the model for a time period spanning a few months. This will lead the model to train with new post covid data. This can be obtained once a vessel arrives at the port, and real delay can be calculated as it was done in methodology section 3.4. This again will be used to train models repeatedly when in deployment. After the testing period, the model will be judged upon the real value it proved to the port of Southampton.

Moreover, as the data regarding the model is publicly available, other stakeholders can easily implement it.

However, as the model has some inaccuracy, such as Ontime Recall is 69% meaning 31% of the time when it predicts Late when the vessel is actually Ontime, this can cause many issues for port or other stakeholders as they will not be prepared for this vessel. The model needs to be further improved by getting more data and relevant data, such as weather and AIS information, as they have been shown to improve ETA predicting as per (Meijer, 2017; Parolas, 2016) research which was discussed in section 2.3 of Literature Review. Also, data such as the previous port and the distance from the previous port can be important.

Chapter 6 - Conclusion

*This **Chapter** concludes the finding of the whole project and provides insight on further research. **Section 6.1** describes the project's key findings along with answering **R01, R02, R03, R04**. **Section 6.2** Identify the limitations of the project, and last **Section 6.3** suggest further possible research based on this project.*

6.1 Main Findings

This project's primary objective was to create an ETA prediction model for Stakeholders involved in container ship transport with a focus on Port of Southampton. These Stakeholders includes Vessels Operators, Terminal and Seaport Operator, Hinterland Transportation and Importers. Their activities and planning depend on the accurate ETA of vessels, as was discussed in Section 2.1. This also answers R01- *What benefit will precise forecast of containership arrival times at container terminals provide for port planning authorities and other stakeholders involved in container transport?*

Therefore to improve the efficiency of stakeholder factors affecting ETA were identified by literature review. These include: ETA provided by ship's Captain, Weekdays, Weather Data, AIS Data, Service of the Vessel. Thus, 4 Machine learning models were created (Decision Tree, Random Forest, Support Vector Machine, Neural Network) based on ETA of Vessel, weekday, Service of vessel data, vessel characteristic and weekly new cases covid data in the UK. AIS and Weather data were omitted due to lack of availability.

As per the testing result, Random forest performed best and revealed factors that affect the delay of the vessel. These includes: ETA of Vessel, Weekday, Size of the vessel, Vessel Owner or Operator, Days before ETA information was received, Next Port, and most interesting was new weekly covid cases. Some of the factors align with what was found

in literature review Section 2.3. However, this project provides new factors that influence the delay of vessels, such as the size of the vessel and Covid cases. These findings provide answers to R02, R03, R04.

R02 - *What are the primary factors influencing vessel delays?*

- ETA of Vessel
- Size of Vessel
- New Covid Cases
- Next Port
- Day before ETA information is received
- Owner or Operator of vessel

R03 - *Which machine learning algorithm produces the most accurate results?*

answer - Random forest.

R04 - *Is Vessel ETA affected by Covid cases?*

answer - Yes

The Random forest model with an accuracy of 80% can benefit various stakeholders of container transport as their activities are dependent on it. Based on the Late or On-time prediction status of the vessel, stakeholders can prioritise and plan their activities, thus cutting cost and improving efficacy, which can lead to overall optimisation of the container supply chain process.

6.2 Limitation

Model Accuracy is around 80%, and recall of OnTime is only 69%, meaning when a vessel is actually OnTime and model predict OnTime 69%. Thus having 31% were predicted Late but were actually on time, Which can be a problem for Stakeholders as their activities depend on it.

Limited data availability was the main limiting factor for improving model accuracy. Data such as AIS contains current speed, average speed, distance to cover, and other important information. Weather data is also important as weather includes the tide and sea waves where vessel cruise though can have a significant impact (Parolas, 2016).

Moreover, only 876 rows of data were available, which covered only a year. More data in a machine learning model can significantly improve the accuracy of the model (Aggarwal, 2018), thus providing much value to the end-user of the model, i.e., a stakeholder in our case.

6.3 Areas for further Research

This project can provide starting point for further research in this field, Some will be highlighted here:

- As this project establishes a relation between vessel delay and covid cases, it does not provide a scientifically-backed reason, only a hypothesis that needs to be tested and validated.
- Future ETA prediction for vessel projects can include the Turnaround time of the previous port as an input for predicted ETA at the subsequent port. As the delay in the previous port can cause a delay in subsequent port (Stepec et al., 2020)

- Another area for further research can include how port management can integrate the output of the ETA prediction model for resource management.

Reference List

Aggarwal, C. C. (2018) Neural Networks and Deep Learning a Textbook. Cham Springer.

Banerjee, C., Mukherjee, T. and Pasiliao, E. (2019) “An Empirical Study on Generalizations of the ReLU Activation Function,” in Proceedings of the 2019 ACM Southeast Conference. Kennesaw, GA, USA: Association for Computing Machinery, pp. 164–167. doi: 10.1145/3299815.3314450.

Daoud, J. (2017) “Multicollinearity and Regression Analysis,” Journal of Physics: Conference Series, 949(1), p. 012009. doi: 10.1088/1742-6596/949/1/012009.

De Jong, G. and Tavasszy, L. (2014) Modelling Freight Transport. Amsterdam: Elsevier.

Ding, S., Hua, X. and Yu, J. (2014) “An Overview on Nonparallel Hyperplane Support Vector Machine Algorithms,” Neural Computing and Applications, 25(5), pp. 975–982. doi: 10.1007/s0052101315246.

Drewry (2006) The Drewry Container Shipper Insight - Fourth Quarter. London: Drewry Shipping Consultants.

Fancello, G. et al. (2011) “Prediction of Arrival Times and Human Resources Allocation for Container Terminal,” Maritime Economics Logistics, 13(2), pp. 142–173. doi: 10.1057/mel.2011.3.

Flapper, E. (2020) ETA Prediction for Vessels Using Machine Learning, essay.utwente.nl. MSc Thesis. Available at: <http://purl.utwente.nl/essays/82201>.

Gómez, R., Camarero, A. and Molina, R. (2016) “Development of a Vessel-Performance Forecasting System: Methodological Framework and Case Study,” *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 142(2), p. 04015016. doi: 10.1061/(asce)ww.1943-5460.0000316.

Grunow, M., Günther, H. and Lehmann, M. (2007) “Strategies for Dispatching AGVs at Automated Seaport Container Terminals,” in *Container Terminals and Cargo Systems: Design, Operations Management, and Logistics Control Issues*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 155–178. doi: 10.1007/9783540495505_8.

IBM Cloud Education (2020) What Are Neural Networks?, www.ibm.com. Available at: <https://www.ibm.com/cloud/learn/neural-networks>.

Jo, J.-M. (2019) “Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance,” *The Journal of the Korea Institute of Electronic Communication Sciences*, 14(3), pp. 547–552. doi: 10.13067/JKIECS.2019.14.3.547.

Knowles, R., Shaw, J. and Docherty, I. (2008) *Transport Geographies: Mobilities, Flows and Spaces*. Malden, Ma: Blackwell Pub.

Koelpin, G. (2020) *Social Network for Programmers and Developers*, morioh.com. Available at: <https://morioh.com/p/811a5d22bbca> (Accessed: 21 November 2021).

Latorre, M. C., Olekseyuk, Z. and Yonezawa, H. (2020) “Trade and Foreign Direct Investment-related Impacts of Brexit,” *World Econ*, 43(1), pp. 2–32. doi: <https://doi.org/10.1111/twec.12859>.

Lee, H. et al. (2018) “A Decision Support System for Vessel Speed Decision in Maritime Logistics Using Weather Archive Big Data,” *Computers Operations Research*, 98, pp. 330–342. doi: 10.1016/j.cor.2017.06.005.

Lee, H. L., Padmanabhan, V. and Whang, S. (1997) “Information Distortion in a Supply Chain: the Bullwhip Effect,” *Management Science*, 43(4), pp. 546–558. doi: 10.1287/mnsc.43.4.546.

Lee, T., Ullah, A. and Wang, R. (2020) “Bootstrap Aggregating and Random Forest,” in *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*. Cham: Springer International Publishing, pp. 389–429. doi: 10.1007/9783030311506_13.

Lee, W.-M. (2019) *Python Machine Learning*. Indianapolis, In: Wiley.

Li, H. et al. (2017) “Capacity Planning for Mega Container Terminals with multi-objective and multi-fidelity Simulation Optimization,” *IIE Transactions*, 49(9), pp. 849–862. doi: 10.1080/24725854.2017.1318229.

Marcot, B. G. and Hanea, A. M. (2020) “What Is an Optimal Value of K in k-fold cross-validation in Discrete Bayesian Network analysis?,” *Computational Statistics*, 36. doi: 10.1007/s00180-020-00999-9.

Meijer, R. (2017) *ETA prediction: Predicting the ETA of a Container Vessel Based on Route Identification Using AIS Data*, repository.tudelft.nl. MSc Thesis. Available at: <http://resolver.tudelft.nl/uuid:cba0ef59-dd23-49aa-91d5-bed239e27395> (Accessed: 20 November 2021).

Menger, I. (2016) Information Exchange between Deep Sea Container Terminals and Hinterland Parties, repository.tudelft.nl. MSc Thesis. Available at: <http://resolver.tudelft.nl/uuid:df65f8c2-3c27-43ce-b9a3-768d964eef51> (Accessed: 10 November 2021).

Montwill, A. (2014) “The Role of Seaports as Logistics Centers in the Modelling of the Sustainable System for Distribution of Goods in Urban Areas,” *Procedia - Social and Behavioral Sciences*, 151, pp. 257–265. doi: 10.1016/j.sbspro.2014.10.024.

Murugan, P. (2017) “Feed Forward and Backward Run in Deep Convolution Neural Network,” in *International Conference on Image Analysis and Processing. 20th International Conference on Computer Vision and Image Processing*, Trento, Italy, p. 20. Available at: <https://arxiv.org/abs/1711.03278>.

Nargesian, F. et al. (2017) “Learning Feature Engineering for Classification,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne, Australia, pp. 2529–2535.

OECD (2019) OECD Ocean, www.oecd.org. Available at: <https://www.oecd.org/ocean/topics/ocean-shipping/>.

Oshiro, T. M., Perez, P. S. and Baranauskas, J. A. (2012) “How Many Trees in a Random Forest?,” in Perner, P. (ed.) *Machine Learning and Data Mining in Pattern Recognition. 8th International Conference*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 154–168.

Pani, C. et al. (2014) “A Data Mining Approach to Forecast Late Arrivals in a Transshipment Container Terminal,” *Transport*, 29(2), pp. 175–184. doi: 10.3846/16484142.2014.930714.

Pani, C. et al. (2015) “Prediction of late/early Arrivals in Container Terminals – a Qualitative Approach,” *European Journal of Transport and Infrastructure Research*, 15(4). doi: 10.18757/ejtir.2015.15.4.3096.

Parolas, I. (2016) ETA Prediction for Containerships at the Port of Rotterdam Using Machine Learning Techniques, repository.tudelft.nl. MSc Thesis. Available at: <http://resolver.tudelft.nl/uuid:9e95d11f-35ba-4a12-8b34-d137c0a4261d>.

Robinson, R. (2006) “Port-Oriented Landside Logistics in Australian Ports: a Strategic Framework,” *Maritime Economics Logistics*, 8(1), pp. 40–59. doi: 10.1057/palgrave.mel.9100149.

Rodríguez, P. et al. (2018) “Beyond one-hot encoding: Lower Dimensional Target Embedding,” *Image and Vision Computing*, 75, pp. 21–31. doi: 10.1016/j.imavis.2018.04.004.

Scikit-learn (2018) Support Vector Machines — scikit-learn 0.20.3 Documentation, Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/svm.html>.

Sea-Intelligence (2018) Global Liner Performance August 2018 Report, Global Liner Performance August 2018 Report. Available at: https://www.sea-intelligence.com/images/products/glp_84.pdf.

Steinbach, M., Tan, P.-N. and Kumar, V. (2005) *Data mining*. Harlow: Addison-Wesley.

Stepec, D. et al. (2020) “Machine Learning Based System for Vessel Turnaround Time Prediction,” in 2020 21st IEEE International Conference on Mobile Data Management (MDM). doi: 10.1109/mdm48529.2020.00060.

Stopford, M. (2009) *Maritime Economics*. 3rd edn. London, New York: Routledge.

Tran, N. K. and Lam, J. S. L. (2021) “Effects of Container Ship Speed on CO2 emission, Cargo Lead Time and Supply Chain Costs,” *Research in Transportation Business Management*, p. 100723. doi: 10.1016/j.rtbm.2021.100723.

Umair Shafique and Haseeb Qaiser (2014) “A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA),” *International Journal of Innovation and Scientific Research*, 12, pp. 217–222.

UNCTAD (2020) *Trade and Development Report 2020* | UNCTAD, unctad.org. Available at: <https://unctad.org/webflyer/trade-and-development-report-2020>.

Visa, G. P. and Salembier, P. (2014) “Precision Recall Classification Evaluation Framework: Application to Depth Estimation on Single Images,” in Fleet, D. et al. (eds) *Computer Vision – ECCV 2014*. 13th European Conference, Zurich, Switzerland: Springer International Publishing, pp. 648–662.

Wang, T.-F. and Cullinane, K. (2006) “The Efficiency of European Container Terminals and Implications for Supply Chain Management,” *Maritime Economics Logistics*, 8(1), pp. 82–99. doi: 10.1057/palgrave.mel.9100151.

Wang, Z., Liang, M. and Delahaye, D. (2018) “A Hybrid Machine Learning Model for short-term Estimated Time of Arrival Prediction in Terminal Manoeuvring Area,” *Transportation research. Part C, Emerging Technologies*, 95, pp. 280–294. doi: 10.1016/j.trc.2018.07.019.

Wei, X. et al. (2020) “Tugboat Scheduling for Container Ports,” *Transportation Research Part E: Logistics and Transportation Review*, 142, p. 102071. doi: 10.1016/j.tre.2020.102071.

Wirth, R. (2000) “CRISP-DM: Towards a Standard Process Model for Data Mining,” in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29–39.

Xu, J., Zhang, Y. and Miao, D. (2020) “Three-way Confusion Matrix for classification: a Measure Driven View,” *Information Sciences*, 507, pp. 772–794. doi: 10.1016/j.ins.2019.06.064.

Xu, W., Song, D. and Roe, M. (2011) “Production and Raw Material Ordering Management for a Manufacturing Supply Chain with Uncertainties,” in *2011 IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 747–751. doi: 10.1109/IEEM.2011.6118016.

Yu, B. et al. (2010) “Hybrid Model for Prediction of Bus Arrival Times at next Station,” *Journal of Advanced Transportation*, 44(3), pp. 193–204. doi: 10.1002/atr.136.

Appendix

Github Link for the whole Project : https://github.com/MSalman5230/EBUS621_MSc_Project