UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

Master Thesis

on the topic of

Modelling and optimization of ship's fuel consumption using Random Forest Regression (RFR)

Submitted to the Faculty of Engineering
of University Duisburg Essen

by

Hibatul Wafi
3021919

Betreuer:	M. T. Muhammad Fakhruriza Pradana
1. Gutachter:	Prof. Dr.-Ing. B. Noche
2. Gutachter:	Dr.-Ing. Alexander Goudz
Studiengang:	ISE General Mechanical Engineering
Studiensemester:	Summer semester 2023
Datum:	04.05.2023

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Marine industry stakeholders are actively pursuing research on efficient ship operation. This research direction is motivated by the increasing price of fuel oil and stricter environmental regulations. The fuel aboard a ship is referred to as "bunkers" and accounts for a substantial portion of the vessel's operational expenses (OPEX). It is known that bunker fuel takes up more than 50% of voyage costs and constitutes up to 75% of the ship's total operating cost. It can be inferred that energy efficient ship operations that could reduce fuel consumption translate to an increase in profitability (**Stopford, 2009**; **Ronen, 2011**; **Bialystocki and Konovessis, 2016**). Furthermore, efficient operation also means reducing Greenhouse Gas Emissions (GHG). The most recent report by International Maritime Organisation indicated that GHG emissions from shipping make up 2.51% of global emissions (**IMO, 2020**). This alignment in motivation implies that through energy efficient ship operation, marine industry stakeholders gain economic benefits while adhering to stringent environmental regulations.

With that, maritime industry stakeholder actively searches for methods to ensure energy efficient operation. Two approaches are considered, namely technical solutions and operational solutions. Technical solutions involve modification to the vessel's structure and power system. But these solutions are expensive, and it requires engineering innovations (**Yan et al., 2021**; **Li et al., 2022**). Because of this, stakeholders look for cheaper solutions to achieve energy efficient operation. The answer for an inexpensive approach lies in optimisation of operational measures, it carries less cost, and it does not require initial investments. Several recommended solutions can be found in Ship Energy Efficiency Management Plan (SEEMP).

However, greater focus will be given in this thesis towards optimising ship speed as reduction of ship speed has the greatest impact on fuel consumption. Different studies indicated that fuel consumption is correlated through a third-order, non-linear function of the ship speed (**Wang and Meng, 2012**; **Ronen, 2011**; **Du et al., 2019**). The significant impact of ship speed on fuel consumption is further supplemented by reports and studies stating that reducing ship speed by about $2-3$ knots could halve the operating cost of shipping companies (**Stopford, 2009**; **Wijnolst et al., 2009**). For these reasons, slow steaming is the measure that is most widely adopted

by shipping operator.

While inexpensive, optimising operational measures is not an easy and trivial task. Several factors ranging from vessel operational performance to varying weather conditions make it challenging to model the ship speed. Some fuel consumption models, which are based on historical data and ship parameters, lack generalisation capabilities, and it is sensitive towards noisy data. To address this problem, recent research turns towards data-driven approach i.e. machine learning approach to predict ship speed and fuel consumption. These studies reported success in their modelling, citing good generalisation capability and low prediction errors. Despite these successes, maritime experts find it difficult to accept models based on data driven approach, as some data-driven models are complex as well as unintuitive and in some cases can violate basic physical knowledge of the vessel. The performance of the data-driven model is also greatly dependent on both data quantity and quality **(Yan et al., 2021; Gkerekos et al., 2019)**.

As such, prompted by volatility and ever-increasing bunker fuel price, developing a model that could accurately predict Fuel Oil consumption (FOC) could prove to be useful to maritime industry stakeholders. As stakeholders could make critical economical decisions at the most opportune moment without violating the stringent environmental regulations.

1.1 Thesis Objective

This thesis proposes an intuitive, data-driven modelling approach that considers varying ship state and environment conditions to predict fuel consumption. To ensure the abundance of data during modelling, this thesis utilise data fused between Automatic Identification System (AIS) and weather data.

To achieve this, Grey Box Model (GBM) approach is selected. Machine learning approach using tree-based regressor is considered to provide a certain degree of intuitiveness to predict ship over ground (SOG) over different journey periods using fused AIS and weather data. Predicted SOG is then converted to actual ship speed i.e. Speed Through Water (STW). STW will be used as the input for modelling of Fuel Oil Consumption (FOC), which is carried out through Holtrop-Mennen estimation method **(Holtrop and Mennen, 1978, 1982; Holtrop, 1984)**, a power estimation method based on hydrodynamic laws which consider resistance forces exerted by environmental conditions.

The following Research Questions (RQs) could be raised during the development of the model :

- **RQ1**: What are the steps that should be taken to optimise the predictive performance of the model?

- **RQ2**: Is it feasible to fuse AIS data and meteorological data to accurately predict the ship's SOG and subsequently FOC of the ship?
- **RQ3**: Which approximations and empirical equations are suitable to estimate the resistance forces required to estimate the power required by the ship?

1.2 Thesis Boundaries

The following research boundaries are set throughout this thesis:

- Due to the continuous nature of the SOG, only the regression aspect of Random Forest (RF) will be considered.
- The focus of this work is a detailed study of the performance and possible optimisation configuration of different tree-based predictors for SOG. As such, an exhaustive comparison study between different types of machine learning models will not be performed.
- In the case study, the approximation for ship parameters and dimensions is based on a similar type of ship with nearly identical dimensions.

1.3 Thesis Contributions

The GBM approach using the fusion of AIS data and weather data provides the following contributions :

- Economical and independent data source.
- Robust modelling approach that requires minimal data pre-processing and minimal model configuration.
- Comprehensible model that adheres to physical principles and hydrodynamic laws of the vessel.

1.4 Thesis Structure

The thesis is organised with the following structure:

Chapter 1 introduces the problem statement and described the objective and boundaries of the thesis. The novelty of this thesis is declared in this chapter.

Chapter 2 The fundamental aspects of the methodologies used to develop the model will be explained in this chapter.Section 2.1 including literature review of relevant past and present research. The fundamentals of the tree-based model will be discussed in Section 2.2, basic explanation of the parameters used in AIS and weather

data will be given in Section 2.3 and Section 2.4. Section 2.5.1 presents the empirical formulas and parameters used to estimate fuel consumption used by the ship based on various literature studies.

Chapter 3 discuss the methodology used to develop tree-based model used for SOG prediction. The discussion comprises analysis of training data, feature selection and reduction and selection of tuning parameters of the model. The methodology to estimate resistance for ship power estimation will be discussed in this chapter as well.

Chapter 4, the GBM model will be evaluated using appropriate performance metrics and their effectiveness will be discussed. The review of the strength and limitations concerning the GBM method will be discussed here.

Chapter 5 The summary of this study and reflections on the research process will be presented here.

Chapter 2

Theoretical Background

2.1 Literature Review

The literature review in Section 2.1 presents past and present research on utilisation of machine learning methods to achieve energy efficient operation. The concept of different modelling approaches for ship operation will be discussed in Section 2.1.1. Short summary of the data source in the modelling of FOC is given in Section 2.1.2. The review of popular machine learning model used to predict FOC is presented in Section 2.1.3. The performance of tree-based model, which include random forest and extra trees in various research will be discussed in Section 2.1.4. Brief summary of the literature review is presented in Section 2.1.6.

2.1.1 Modelling Approach for Ship Operation

According to **Haranen et al. (2016)** and **Coraddu et al. (2017)**, the modelling strategies to predict fuel consumption are classified into three categories:

White Box Models (WBM)

Based on *a priori* mechanistic knowledge and physical principles of the vessel's system. This means that the dimensions of the vessel's structure, design parameters, and propulsion plant configuration are known.

Black Box Models (BBM)

Purely data driven, and it is developed using data from different sailing journey and historical observations. Contrary to WBM, this approach does not require detailed information on the vessel. This modelling approach can be further split into two categories. *Statistical Modelling* aims to find explanations for relationships between fuel consumption and different factors that affect it. *Machine Learning (ML) Modelling* focuses on the predictive capabilities of the model that could predict fuel consumption

at different points in time.

Grey Box Models (GBM)

Fuse WBM and BBM into a single model that considers both *a priori* knowledge of the vessel and historical sailing data, This method aims to complement the performance of WBM and BBM.

Each of these strategies possesses its strength and limitations. WBMs are developed based on physical and hydrodynamics laws as well as theories of naval architecture, it is transparent and comprehensible, making them the preferred model used by various shipping industries. However, the deterministic nature of WBMs causes them to have poor suitability and generalisability. This is mainly caused due to limited *a priori* knowledge of different vessel dimensions, parameters, and narrow application limits of principle dimensions and form parameters of the vessel. Subsequently, the inability of WBMs to add randomness makes it rigid and restrictive. **(Haranen et al., 2016; Yan et al., 2021)**

BBMs in general have a good fitting ability for training data and good predictive accuracy for unseen data. BBMs developed using machine learning approach can generalise better compared to BBMs that are based on statistical modelling **(Petersen et al., 2012a)**. BBMs are purely data driven, which means BBMs do not require former knowledge of vessel principle dimensions and form parameters. With increasing amount of data, better generalisation performance and handling of noisy data should be expected in a BBM. However, for the same reason, the quality of BBM model is highly dependent on data quantity and quality. For BBMs based machine learning approach, the amount of data is a major factor in determining the effectiveness of machine learning **(Halevy et al., 2009)**. Data driven approach means that BBMs neglect basic vessel physical knowledge and are generally complex making it challenging to analyse and explain. For these reasons, experts in shipping industries are critical of models that do not include basic vessel knowledge and those that violate concepts of the domain knowledge in serious ways **(Yan et al., 2021)**.

Hence, GBMs are introduced to address the limitations of both WBMs and BBMs by combining the mechanistic knowledge of the ship and physical principles of the vessel's system with BBM models, which possess good predictive capability. Despite these advantages, **Yan et al. (2021)** noted that GBM approach is not a common approach, recent research to predict fuel consumption are mainly dominated by BBM approach, specifically BBM based on machine learning approach.

2.1.2 Review of data source used for FOC model

The modelling of FOC using GBM requires both components of WBM and BBM. For the BBM modelling part using machine learning approach, it is especially important

to ensure sufficient amount of good quality data to be available for model training to ensure precise and accurate training of the model (Halevy et al., 2009). It summarised by Yan et al. (2021), that the modelling of FOC use the following types of data source:

(Daily) Noon Report

Daily reports manually filed by ship's chief engineer and sent by the ship's masters to the shipping company and shore management. The reports include informations on types of daily fuel consumption, basic voyage information (e.g. ship location, load condition), sailing behaviour information e.g. (average sailing speed, average engine revolution per minute (RPM)), as well as sea and weather conditions. While it provides relevant information regarding the ship operation, the inherent problem of daily and manual data entry means that the quality and quantity of data cannot be guaranteed.

Sensor Data

Data obtained from installed sensor onboard the vessel. This may include fuel flow sensors, Global Positioning System (GPS) receiver and wind speed sensors are among the possible sensors that can be installed onboard a vessel. Sensor data address the issues of data quantity from noon report, as pointed out in the study by Gkerekos et al. (2019) for the prediction of daily FOC. The machine learning models, which are produced by the Automated Data Logging and Monitoring (ADLM) system outperforms the models that used noon data for their training by $5-7\%$ for a collection period of 3 months of the ADLM system and 2.5 years for the noon data. However, installing onboard sensors may be complex and costly (Petersen, 2011) and the resulting sensor data will need to be handled properly to account for error in the sensors.

AIS Data

Apart from its intended use as collision avoidance system, AIS data have seen potential usage in ship behaviour analysis and environmental analysis. The Green House Gas (GHG) study by IMO (IMO, 2020; Smith et al., 2015), uses AIS to estimate global shipping emission inventories. Rakke (2016) proposed a methodology termed ECAIS to calculate ship emissions based on the fuel consumption from AIS data. Through Holtrop-Mennen approximation and literature approximation, the ship's power propulsion can be determined which is subsequently used to predict specific fuel consumption. Kim et al. (2020) used publicly accessible AIS data, ship static data, and environmental data to estimate EEOI using big data technology. Generally, the study using AIS data is done to achieve independence from the need to use commercial databases. The details of AIS data will be discussed in Section 2.3

2.1.3 Review of ML approach to predict FOC

Modelling of FOC using *machine learning* approach generally focus on prediction of unseen data. The general framework usually include collection and preprocessing of ship operational data, training and validation of the model, and evaluation and selection of the most appropriate model. Some machine learning models allow further hyperparameter tuning of the model and in case of data rich environment, the data can be further split into test data to further validate the performance of machine learning model.

The study by **Yan et al. (2021)** indicated that the majority of recent research that uses machine learning approach employ ANN as the model to predict FOC. ANN models are powerful models capable of modelling nonlinear data which are based on theories on how the brain works. The outcome is modelled by intermediate set of unobserved variables known as hidden layer. **(Kuhn and Johnson, 2013)**. Back propagation neural networks, Multi Level Perceptron (MLP), and wavelet neural networks are some examples of ANN model subclasses.

ANN has shown respectable performance in its attempt to predict FOC. **Petersen et al. (2012b)** reported Root Mean Square Error of 47.2 L/h for the fuel flow i.e. FOC. To put this into context, the fuel flow in their case study fluctuates between $1000-2500$ L/h. **Bal Beşikçi et al. (2016)** considered sailing speed, trim, wind, sea effects, propeller pitch, and engine rotation speed as input variables to predict FOC per hour and achieved model fit score of $R^2 = 0.759$ in test set. Other studies also reported similar range of results using ANNs **(Yan et al., 2021)**.

However, the development of ANN models is a challenging task. ANN models tend to overfit when there is shortage of data, as such, regularisation is necessary to improve model performance. The balancing process during regularisation is a demanding task and unsuitable regularisation may lead to counterintuitive prediction results. Adding layers is computationally expensive, and it does not always guarantee promising results **(Hastie et al., 2009)**. Additionally, in machine learning terms, ANN is classified as a black box model, which makes it unintuitive and lacking in interpretability **(Géron, 2019)**, this particular limitation cause shipping industry expert generally reluctant to accept the model generated using machine learning approach.

2.1.4 Tree-Based Model as FOC model

Concerning interpretability, modelling approaches such Linear Regression (LR), KNN and tree-based models have shown superior interpretability in comparison to ANNs. LR can explain the effect of each input variable on the output through the coefficients. KNN searches for the nearest neighbour and their closeness is evaluated through distance measurement algorithms such as Euclidean distance. Additionally, LRs and KNNs also offer easy implementation and adequate explainability. However,

both approaches suffer from sensitivity to outliers and noise in data.

This brings us to tree-based model, a supervised, highly interpretable machine learning modelling approach capable of performing classification tasks for discrete data and regression tasks for continuous data. According to summary of **Yan et al. (2021)**, it is not as popular as ANN, however some literature work and studies have indicated its benefits and performance superiority over other machine learning modelling approaches:

Soner et al. (2018) used the ferry dataset from **Petersen et al. (2012b)** to predict FOC using tree-based model, which includes bagging, random forest (RF), and bootstrap. From the test dataset, the random forest model achieved RMSE of 43.5 L/h for the fuel consumption. Which suggested improvement from ANN model from the study of **Petersen et al. (2012b)**.

Yan et al. (2020) used random forest (RF) model to predict FOC for a voyage of a dry bulk ship using ship operational data i.e. ship noon data and sea and weather data from noon report and EMCWF. The prediction model considered ship sailing speed, total cargo weight and meteorological conditions and RF model obtained mean absolute percentage error (MAPE) of 7.91% for the FOC. The RF model displayed superior result in comparison to Decision Tree Regressor (DTR), ANN, LASSO, and SVR.

The advantage of tree-based model is further highlighted by **Gkerekos et al. (2019)**. The study compared the performance of different machine learning models to predict ship's FOC per day using both noon data and automated data logging and monitoring (ADLM) system from a bulk carrier. This research concludes that tree-based model displayed good prediction performance on both noon data and sensor-based data. ETR achieved remarkable model fit score of 89% using the noon data and 97% when using the data from ADLM system, outperforming ANN, SVR, and RFR models.

Li et al. (2022) performed more extensive research on the effects of data fusions between meteorological data, ship voyage data, and AIS data on different machine learning models to predict the ship's FOC. The study classified ETR and RFR as tree-based model which is produced by *bagging ensemble strategy*. While AdaBoost (AB), Gradient Tree Boosting (GB), XGBoost(XG) and LightGBM (LB) are classified as tree-based models produced by *boosting ensemble strategy*. The study recommends all tree-based models that are produced by *boosting ensemble strategy* and ETR to be used to model energy efficient operation. Additionally, RFR shows the best robustness among the proposed model in the study.

Abebe et al. (2020) attempted to use machine learning approach to predict SOG of the ship. In this study, AIS data and noon-report weather data from 14 tracks and 62 ships are used for model training. The model considered the ship draught, ship dynamic information, tonnage, and environmental conditions. The result of this study exhibited the feasibility of using AIS data and meteorological data to predict SOG of

the ship. The results also further indicated the strength of tree-based model, on test dataset, ETR achieved the best result with model fit of 98.47% and RMSE of $0,234$ knots. It is also reported that ETR achieved better performance with about half of the computational cost of RFR.

2.1.5 Review of WBM for FOC prediction

To predict the FOC of a ship, WBMs usually calculate the resistances encountered by the vessel based on physics and hydrodynamic laws. The total resistance is summed from resistance of calm water resistances and additional resistance from wind, wave, and other external factors. The corresponding engine power at a particular speed can be calculated, and consequently the FOC can be calculated.**(Haranen et al., 2016)**

The methods from **Guldhammer and Harvald (1974)**, **Hollenbach (1999)**, and **Kristensen and Lützen (2012)** use different formulations, assumptions, and input variables for engine power estimation. For this thesis, the main focus will be the use of the estimation method from Holtrop-Mennen **(Holtrop and Mennen, 1978, 1982; Holtrop, 1984)**. Holtrop-Mennen estimate method allowable application range is suitable in most of the cases. This is indicated by studies from **Rakke (2016)** and **Kim et al. (2020)**. **Rakke (2016)** used ship operational and mechanical data from various works of literature and AIS data for input variables to estimate the engine power using Holtrop-Mennen method to subsequently calculate FOC. The FOC is then used to estimate GHG emissions for different ships and the study reported about 5% error rate during model testing. **Kim et al. (2020)** successfully estimated Energy Efficiency Operational Index (EEOI) without actual FOC. The study used AIS data as well as publicly accessible weather data and ship static information. The approach in this study used Holtrop-Mennen method to estimate engine power which is consequently used to calculate FOC for EEOI estimations.

2.1.6 Conclusion of Literature Review

As termed by **Yan et al. (2021)**, the GBM model in this thesis falls under the category of sequential GBM, where the BBM and the WBM will be developed in series and combined to form a single GBM. The BBM will be developed to perform initial prediction and the resulting prediction will be passed into the WBM. The use of tree-based regressor, which will be used to predict SOG, provides solution to the problem of poor interpretability of some machine learning models. Furthermore, tree-based models can outperform most of the available machine learning models while providing added benefits of little requirement for data preprocessing and relatively cheap computational cost. The selection of Holtrop-Mennen as engine power estimation method is justified by the application range of the methodology and successes from previous studies.

2.2 Tree-based model

Decision Tree, Random Forest and Extra-Tree are classified as tree-based model, which is supervised machine learning model capable of classification tasks for discrete variables and regression tasks for continuous variables. In this section, the theory of Decision Tree (DT), Random Forest (RF) and Extra Tree (ET) will be discussed in detail in Section 2.2.1, Section 2.2.2 and Section 2.2.3.

2.2.1 Decision Tree

The principle of decision tree as a predictor can be defined as one or more nested `if-then` statements based on a rule that partitions the data into partition space as shown in Figure 2.1. Alternatively, the partition space generated from `if-then` statements can be represented using binary tree representation, which is more interpretable as multiple input response can be represented by a single tree.**(Kuhn and Johnson, 2013; Hastie et al., 2009)**

A decision tree consists of the following type of nodes, ***Root node*** defines the topmost node. ***Leaf nodes*** are also termed as terminal nodes, it is the node that will give the final prediction output. The ***Internal Node*** is defined as the nodes between the root node and leaf node. The process of dividing a node into successive nodes is called ***splitting***. The node that is being split is called ***parent node*** and the successive nodes that are created are called ***child nodes***. To grow a tree in a regression task, the splitting process is commonly regulated by Mean Square Error (MSE). The tree growth algorithm are based on Classification and Regression Tree (CART).

To understand the principle of selection for the feature, k_t , of the parent node and splitting rule, t_k , for data partition, the following example will be presented:

For the selection of the optimal splitting rule t_k: Given a case with single feature k and response y with m data points present. The algorithm starts by looking for possible splits between two distinct data points y. This split results in two distinct partition spaces. For each partition space S_1 and S_2, the mean is calculated by dividing the sum of response y with the amount of data points m for each respective partition space S_1 and S_2.

This step is then followed by calculating the sum of squared error (SSE) of each data point in partition space S_1 and S_2 and dividing it by the number of data points m_{s_1} and m_{s_2} respectively to obtain the MSE. Subsequently, the MSE from the respective partition space S_1 and S_2 is summed. The process is then recursively repeated until a threshold t_k that produces minimum sum of MSE is found, this threshold will be selected as splitting rule for the parent node and correspond to the threshold that minimise the cost function $J(k, t_k)$, with $\hat{y}_{S_i}$, being the mean of the response, y_{S_i}, in partition space S_i. **(Géron, 2019; Kuhn and Johnson, 2013)**:

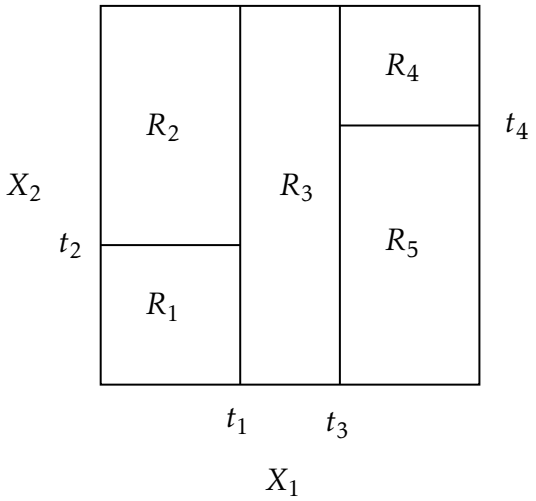

Figure 2.1: Example of partition space (**Hastie et al., 2009**)

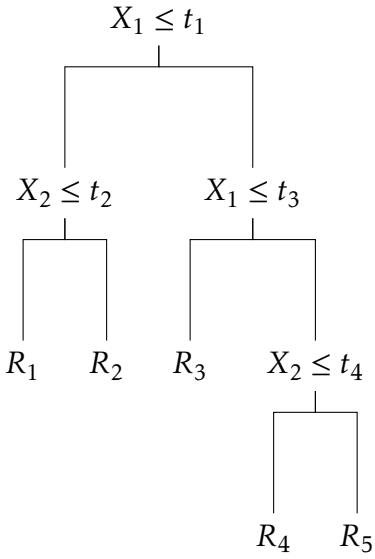

Figure 2.2: Example of partition tree (**Hastie et al., 2009**)

$$\text{MSE}_{S_i} = \frac{1}{m_{S_i}} \text{SSE}_{S_i} \quad \textbf{where} \quad i = (1, 2) \tag{2.2.1}$$

$$J(k, t_k) = \frac{1}{m_{S_1}} \text{SSE}_{S_1} + \frac{1}{m_{S_2}} \text{SSE}_{S_2} \begin{cases} \text{SSE}_{S_i} = \sum\limits_{i \in S_i} (\hat{y}_{S_i} - y_{S_i})^2 \\ \hat{y}_{S_i} = \frac{1}{m_{S_i}} \sum\limits_{i \in S_i} y \end{cases} \tag{2.2.2}$$

For the selection of the most optimal feature for parent node t_k: Similar principle is also applied for the selection of the most optimal feature for the parent node. Consider there are k_t features, then for each respective feature $k_1, k_2, \ldots, k_t$, The MSE for each of the features is calculated following the cost function $J(k, t_k)$. The feature that can best ***minimise*** the cost function will be selected as the root node of the tree. The subsequent selections of the feature for the parent node follow the same principle. (**Hastie et al., 2009**; **Géron, 2019**).

Once complete, then the partition space is further split into two more regions to look for the next possible split that minimise the cost function $J(k, t_k)$. This process is recursively continued until the number of samples to split falls under a certain threshold or when it cannot find a split that can further reduce MSE.

The resulting decisions for the best possible splits can be represented using binary tree, this makes decision tree highly interpretable and easy to implement. The inherent logic structure from `if-then` statements means that it can handle various types of data (sparse, skewed, continuous, categorical, etc.) without the need for data pre-processing. Decision tree implicitly conducts feature selection which is a desirable trait for many modelling problems (**Kuhn and Johnson, 2013**).

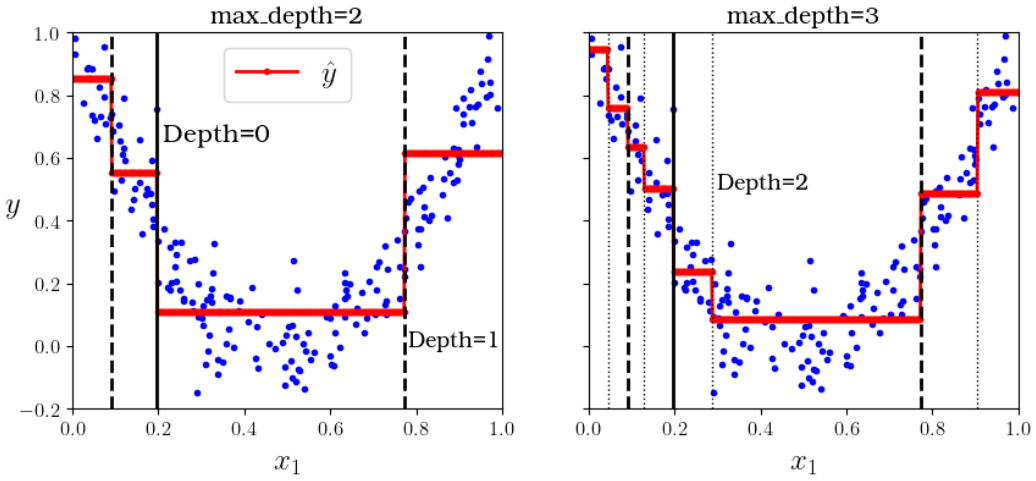

Figure 2.3: Prediction of two Decision tree regression models (Géron, 2019)

However, a single decision tree suffers from overfitting when the model is unconstrained. The logical principle of `if-then` statements means that decision tree will attempt to fit the training data as closely as possible. Furthermore, a single decision tree model tends to be unstable, altering the data will cause drastic changes in the structure of the tree, there exist possibilities where completely different sets of splits might be found resulting in different interpretations (Hastie et al., 2009; Kuhn and Johnson, 2013).

From Figure 2.1, it can be implied that each decision boundaries are orthogonal to an axis i.e. all splits are perpendicular to an axis and this form rectangular subspaces for each predicted value. If the relationship between predictors and response cannot be adequately defined by the rectangular subspaces, then tree based models will suffer from larger prediction error than other kinds of models (Kuhn and Johnson, 2013).

Therefore, it is necessary to regularise i.e., restrict the decision tree's freedom to grow during model training. Overfitting could be reduced by controlling how deep the tree can grow through the `max_depth` parameter. Additionally, setting the amount of minimum number of samples a leaf node has, through `min_samples_leaf` can alleviate overfitting as well, as shown in Figure 2.4. Other regularisation techniques will be discussed in Section 3.3.2.

Regularisation of decision tree will help to address the overfitting issues and improve the robustness of the model, this may result in better generalisation capability. Nonetheless, in order to attain significant improvements in the performance of decision tree model, it is necessary to seek alternative solutions.

2.2.2 Random Forest

Ensemble learning is one of the possible solutions to improve the performance of DT regressors. The main idea of ensemble learning is combining the strengths of a

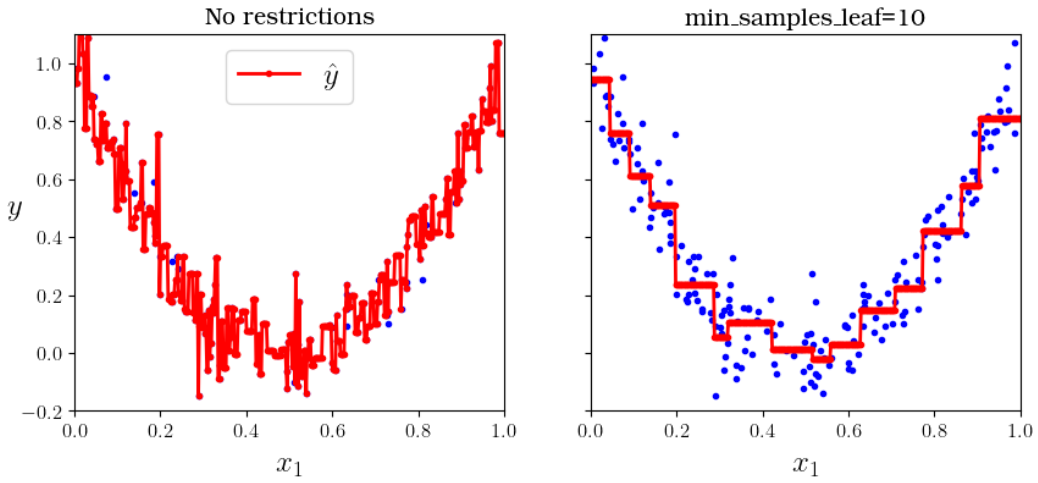

Figure 2.4: Regularising a Decision Tree regressor (Géron, 2019)

collection of simpler base models (Hastie et al., 2009). The algorithm, involving the creation of bootstrap samples, random selection of splitting feature and aggregation of the prediction is termed by Breiman (2001) as ***Random Forest***. It involves combination of multiple learning algorithms, known as weak learners. In random forest, each of these learners are individual decision tree.

The most common ensemble methods are *boosting* and *bagging*. In boosting, the learner evolves over time, where successive trees are dependent on the earlier trees. In bagging (short for *bootstrap aggregating*) each tree is trained using bootstrap sample of the training set i.e. this means that a sample of the training dataset is randomly selected and allowed to appear more than once[1]. Each model in the ensemble then generates a prediction from the bootstrapped sample and the predictions are aggregated across the learners (Tin Kam Ho, 1995; Breiman, 2001).

The performance of bagging can be further improved by reducing correlation between trees i.e. de-correlating trees. This can be achieved by adding randomness during tree construction process. Dietterich (2000) introduced the idea of random split selection, which means that a feature k will be selected from a random subset of feature. From this random subset, the assignment of the feature for the parent node follows the CART algorithm described in Equation (2.2.2). Further randomness is added by exploiting the instability of single decision tree mentioned in Section 2.2.1.

The methodology introduced in random forest address the tendency of decision tree to overfit and the issue of lack of robustness. De-correlating trees means that each learner is independent of each other, and the combination of many independent, strong learners yields an improvement in error rates i.e. reduction in variance and robustness against noisy response. It is also proven by Breiman (2001) that random forest cannot overfit, that means growing more trees should not affect the performance of random forest, albeit with a greater computational burden. Both Kuhn

[1] This sampling technique is referred to as sampling *with* replacement

and Johnson (2013) and Hastie et al. (2009) reported that remarkable prediction results can be obtained without extensive tuning of tree parameter.

However, random forest loses the benefit of interpretability of tree-based model. Due to ensemble nature of random forest, it is not possible to gain an understanding between the feature and the prediction. Nevertheless, it is still possible to quantify the impact of each feature in the ensemble (Kuhn and Johnson, 2013)[2]. Random forest also tends to perform poorly with small number of samples (Hastie et al., 2009). Nevertheless, it is possible to traverse through a single tree to see the path taken to reach the predicted value.

2.2.3 Extra-Trees (Extremely Randomised Trees)

Extra-trees (Extremely Randomised Trees) is introduced by Geurts et al. (2006) to further randomise random forest and further de-correlate the trees in the forest. Unlike random forest, which selects the optimal split by selecting the best feature among randomly selected subset of features, Extra-trees selects a split at random. Extra-trees also does not bootstrap the sample[3] and uses the whole training dataset. The random selection of split means that it saves computational power and the increase in variance caused by tree de-correlation can be countered by increasing the number of trees in the ensemble.

2.3 AIS Data

2.3.1 Overview of AIS

Automatic Identification System (AIS) is an automated tracking system onboard ships to automatically transmit information about the ship to other ships and coastal authorities, AIS was developed to avoid ship collision accidents. As part of the revised new chapter V of SOLAS[4] regulation, International Maritime Organization (IMO) requires all international voyage ships of 300 gross tonnage (GT) and upwards, cargo ships with 500 GT not engaged on international voyage, and all passenger ships irrespective of size to be equipped of AIS class A equipment (Yang et al., 2019; IMO, 2015).

AIS uses Very High Frequency (VHF) with special protocol for communication system for information exchange between the ships. This information will be received by either ships directly, buoys, Land based AIS transceivers (T-AIS) and satellites (S-AIS). The information transmitted by AIS is distinguished into three different types.

[2] Known as feature importances in `Scikit-Learn`
[3] This sampling technique is referred to as sampling *without* replacement
[4] International Convention for the Safety of Lives at Sea

Information Item	Description
Static	
MMSI	MMSI number of vessel
Callsign	Callsign of vessel
Name	Name of the vessel
IMO	IMO number of the vessel
Length	Length of vessel
Width	Width of vessel
Ship Type	Describes the AIS ship type of this vessel
Dynamic	
Ship's position	Automatically updated from position sensor connected to AIS. Longitude and Latitude.
Position time stamp in UTC	Automatically updated from ship's main position sensor. Format: DD/MM/YYYY HH:MM:SS
Course over Ground (COG)	***If available***, automatically updated from ship's main position sensor connected to AIS.
Speed Over Ground (SOG)	***If available***, automatically updated from the position sensor connected to AIS.
Heading	Automatically updated from the ship's heading sensor connected to AIS
Navigational status	Navigational status information has to be manually entered by the Officer on Watch (OOW) and changed as necessary. For example : "*underway by engines*","*engaged in fishing*","*at anchor*".
Rate of Turn (ROT)	***If available***, Automatically updated from the ship's ROT sensor or derived from the gyro.
Voyage Related	
Ship's draught	To be manually entered at the start of the voyage using the maximum draft for the voyage and amended as required
(Hazardous) Cargo Type	Type of cargo from AIS message.
Destination and ETA	To be manually entered at the start of the voyage and kept up to date as necessary.

Table 2.1: Structure of AIS data **(IMO, 2015)**

Static information which is entered into the AIS on installation. **Dynamic information**, which is automatically updated from the ship's sensors connected to AIS and **voyage-related information**, which might need to be manually entered and updated during the voyage. The structure of the AIS data that is relevant to this thesis is summarised in Table 2.1**(IMO, 2015)**.

AIS is also further differentiated by its equipment class. The classification is based on the reporting interval and the type of information that is conveyed. **Class A** autonomously report their position within 2-10 seconds interval, depending on the state of ship's movement. The reporting interval is less frequent at 3 minutes, When the ship is at anchor or moored and moving slower than 3 knots. Class A AIS is also capable of sending safety related information, meteorological and hydrological data, electronic broadcast to mariners and marine safety messages. **Class B** reports at longer interval and at a lower power. They can only receive safety related messages, not send them. **(Rakke, 2016; IMO, 2015)**

It is also stated by **Yang et al. (2019)** that AIS data can be combined with data from other databases to provide additional information such as:

- Port to port average speed, the voyage time can be calculated from the time stamps reported by AIS data; the voyage distance can be found from corresponding navigation distance tables.
- Cargo weight which can be estimated from draught and ship size.
- Technical ship specification from fleet database which can be derived from IMO number.
- Port to port bunker consumption which can be estimated based on the speed, technical ship specification and distance between two ports.

2.3.2 Speed Correction

The speed that is shown in AIS is the speed over ground (SOG). However, the ship actual speed i.e. speed through water (STW) will be required to calculate the bunker fuel consumption. Therefore, the SOG will need to be corrected for STW. This correction is performed by considering the current speed V_c and the direction of the current γ *with respect to True North*. In principle, STW will be greater than SOG, when the current is moving against the current as the ship tries to compensate for the current to maintain the SOG. Whereas, the STW will be greater than the SOG when the current is moving in the same direction of the ship movement.

To calculate the correction, this study will adopt the methodology proposed by Kim et al. **(Kim et al., 2020)** and Yang et al. **(Yang et al., 2020)**. The x and y component of SOG can be obtained through vector decomposition using the ship's heading angle α *with respect to True North*. Similar vector decomposition is also performed for current speed V_{current}, it is resolved with current direction γ *with respect to True North*:

$$V^x_{\text{SOG}} = V_{\text{SOG}} \cdot \sin(\alpha) \tag{2.3.1}$$

$$V^y_{\text{SOG}} = V_{\text{SOG}} \cdot \cos(\alpha) \tag{2.3.2}$$

$$V^x_{\text{current}} = V_{\text{current}} \cdot \sin(\gamma) \tag{2.3.3}$$

$$V^y_{\text{current}} = V_{\text{current}} \cdot \cos(\gamma) \tag{2.3.4}$$

Then the resulting equation to determine STW, including the current compensation, is given by:

$$V^x_{\text{STW}} = V^x_{\text{SOG}} - V^x_{\text{current}} \tag{2.3.5}$$

$$V^y_{\text{STW}} = V^y_{\text{SOG}} - V^y_{\text{current}} \tag{2.3.6}$$

$$V_{\text{STW}} = \sqrt{(V^x_{\text{STW}})^2 + (V^y_{\text{STW}})^2} \tag{2.3.7}$$

2.3.3 Source of error in AIS

Errors and inaccuracies may still exist in AIS data. The main source of errors is caused by data that requires manual entry such as static information and voyage related information which include estimated time of arrival (ETA) and draught. There exist cases where MMSI is shared by different ships even though it is supposed to be unique. The data that is automatically connected by sensors can be erroneous, this may happen when the sensors are faulty or when it is not properly installed (**Yang et al., 2019**). Therefore, data preprocessing of AIS data is an important step to ensure correct representation of the ship state.

2.4 Weather data

During voyage, a vessel may encounter winds and waves from different directions with varying degree of magnitude. This may affect the vessel's path taken during the voyage and also ship performance such as speed and engine power, furthermore it may also affect the seakeeping capability of a vessel (**Molland, 2011**). It is important to consider different weather conditions to ensure accurate and precise estimation of required engine power by the vessel. With that in mind, the discussion in this section will focus on definition of wind and wave effects, as well as the relation between some of these parameters.

2.4.1 Definitions of weather parameters

Wind Waves and Swell

Wind Waves are also known as wind sea, wind waves are irregular and short-crested waves generated by local wind. **Swell** are waves that travel outside the wave generation area and are no longer the result of wind, they take on regular and long-crested appearance (**Holthuijsen, 2007**)

Significant Wave Height, $H_{1/3}$

It is defined as the mean of the highest one-third of waves in the wave record. The distribution of wave heights can be represented by probability density function. Hence, the term "highest one-third of waves" here means the region of wave heights that belong in the upper one-third of a probability density function, this is illustrated in Figure 3.4. From this distribution, the relation between significant wave height $H_{1/3}$, the highest ten percent of waves H_{10}, maximum wave height H_{max} and average wave height $\overline{H}$ can be summarised as follows (**Bretschneider, 1965**; **Holthuijsen, 2007**):

$$\overline{H} = 0.625 \cdot H_{1/3} \tag{2.4.1}$$

$$H_{10} = 2.03 \cdot \overline{H} = 1.27 \cdot H_{1/3} \tag{2.4.2}$$

$$H_{max} = 2 \cdot H_{1/3} \tag{2.4.3}$$

Additionally, **Bitner-Gregersen (2005)** and **Nielsen and Dietz (2020)** described the relation between the significant wave height, wind wave height and swell height through following equation:

$$H_{1/3} = \sqrt{(H_{swell})^2 + (H_{windwave})^2} \tag{2.4.4}$$

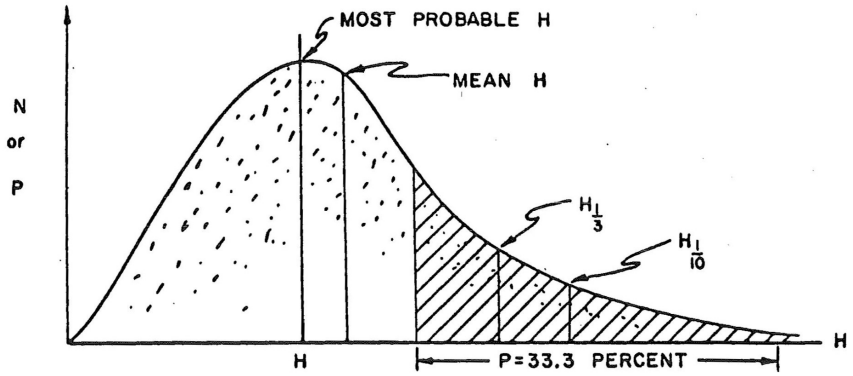

Figure 2.5: Statistical distribution of wave heights **(Bretschneider, 1965)**

Wave Period

Defined as the time interval between the start and the end of a wave. Some characteristics of wave period can be derived to define wave spectrum.

Wave Spectrum

The most important form in which ocean waves are described. Wave spectrum characterises all possible observations of the waves which include wave heights, frequencies i.e. period and wave direction. For example, **Bitner-Gregersen (2005)** stated that the state of the sea can be described through the significant height $H_{1/3}$ and spectral peak T_p with the help of Torsethaugen peak, given average wave spectrum T_f and constant $a_f = 6.6$ **(K. Torsethaugen et al., 2004)**.

$$T_p = a_f \cdot H_{1/3} \tag{2.4.5}$$

$$\text{Sea State (SS)} = \begin{cases} \text{Swell dominated} & \textbf{if} \quad T_p > T_f \\ \text{Wind sea dominated} & \textbf{if} \quad T_p \leqslant T_f \end{cases} \tag{2.4.6}$$

2.5 General concept of ship propulsion

A ship's bunker fuel consumption in actual operating conditions are affected by several factors including the operating parameter of the ship's engine, propeller efficiency and encountered resistance by the ship. Furthermore, a ship's propulsion

power is correlated to the sailing speed (SOG) and meteorological conditions (Lang, 2020). Therefore, in addition to the calm water resistance R_{CALM}, the additional resistance caused by wind R_{WIND} and wave R_{AW} should be considered to estimate the total resistance of the ship R_{TOTAL}. The power needed to propel a ship forward at a given ship STW v_S, to overcome R_{TOTAL} is defined as **effective power** P_e:

$$R_{TOTAL} = R_{TOTAL} + R_{AW} + R_{AA} \tag{2.5.1}$$

$$P_e = R_{TOTAL} \cdot v_S \tag{2.5.2}$$

The effective power P_e is transmitted through the shaft connected to the main engine of the ship which generates power to rotate the propeller of the ship, which is termed as **brake power of the engine,** P_b. The brake power can be calculated through effective power by considering the **shaft efficiency** η_s**, hull efficiency** η_h**, relative rotative efficiency** η_r **and open water efficiency** η_o:

$$P_b = \frac{P_e}{\eta_s \cdot \eta_h \cdot \eta_r \cdot \eta_o} \tag{2.5.3}$$

The bunker fuel consumption then can be calculated by multiplying the brake power P_b with the Specific Fuel Oil Consumption (SFOC) and the operation time:

$$FOC = P_b \cdot SFOC \cdot \mathcal{T}_{operation} \tag{2.5.4}$$

2.5.1 Holtrop & Mennen's Method

This power prediction method was applied in late 1970s and early 1980s by J. Holtrop and G.G.J Mennen and it was based on regression analysis of vast model tests and trial data of MARIN, the model basin in Wageningen, The Netherlands. This gives Holtrop Mennen wide applicability range and the only method that adopted the use of the ITTC form factor k. The resistance in this method are calculated as dimensional force and additionally provides formula to estimate hull-propeller interaction thrust dedudction, full scale wake fraction and relative rotative efficiency (Birk, 2019).

Application Range

The publication from Holtrop and Mennen (1978, 1982); Holtrop (1984) does not provide explicit information regarding the application range of the method. However, from experience of Birk (2019), reasonable estimates from the method can be achieved for the following conditions:

$$\begin{gathered} Fr \leqslant 0.45 \\ 0.55 \leqslant C_p \leqslant 0.85 \\ 3.9 \leqslant \frac{L}{B} \leqslant 9.5 \end{gathered} \tag{2.5.5}$$

Parameter	Symbol	Remarks
Required Parameters		
Length in waterline	L_{WL}	
Molded breadth	B	
Molded mean draught	T	typically $T = \frac{1}{2}(T_A + T_F)$
Molded draught at aft perpendicular	T_A	
Molded draught at forward perpendicular	T_F	
Volumetric displacement (molded)	V	alternatively use the block coefficient as $C_B = V/BTL_{WL}$
Prismatic coefficient (based on $L_W L$)	C_P	
Midship section coefficient	C_M	or use $C_M = C_B/C_P$
Waterplane area coefficient	C_{WP}	$C_{WP} = (1 + 2C_B)/3$, from **Schneekluth and Bertram (1998)**
Longitudinal Centre of buoyancy	ℓ_{CB}	$\ell_{CB} = 0.44Fr_{\text{design}} - 0.094$, from **Guldhammer and Harvald (1974)**
Area of ship and cargo above waterline	A_V	projected in direction of v_S
Immersed transom area	A_T	
Transverse area of bulbous bow	A_{BT}	Measured at forward perpendicular
Height of centre A_{BT} above basis	h_B	has to be smaller than $0.6T_F$
Propeller Diameter	D	
Propeller expanded area ratio	A_E/A_0	
Stern shape parameter	C_{stern}	
Optional Parameters		
Wetted surface (hull)	S	
Wetted Surface of appendages	S_{App}	bilge keels, stabiliser fins, etc.
Half angle of waterline entrance	i_E	
Diameter of bow thruster tunnel	$d_T H$	

Table 2.2: Required and optional input parameters for Holtrop & Mennen's method according to **Birk (2019)**

2.5.1.1 Calm water resistance

The calm water resistance R_{CALM} is broken down into several components and can be approximated using the following relation:

$$R_{CALM} = R_F(1 + k_1) + R_{APP} + R_W + R_B + R_{TR} + R_A \tag{2.5.6}$$

Frictional Resistance R_F

R_F is calculated using the ITTC-1957 frictional resistance correlation line C_F as the basis of a representation of a resistance plate with a wetted surface area S of bare hull.

$$R_F = \frac{1}{2}\rho v_S^2 S C_F \tag{2.5.7}$$

The frictional coefficient C_F can be calculated through the Reynold number Re for a given ship speed v_S and kinematic viscosity ν:

$$C_F = \frac{0.075}{[\log_{10}(Re) - 2]^2} \quad \textbf{where} \quad Re = \frac{v_S L_{WL}}{\nu} \tag{2.5.8}$$

If not known, then the wetted surface area of bare hull S can be estimated by the following formula:

$$S = c_{23} L_{WL}(2T + B)\sqrt{C_M} + 2.38\frac{A_{BT}}{C_B} \tag{2.5.9}$$

with the factor c_{23} given as :

$$c_{23} = \left[0.453 + 0.4425C_B - 0.2862C_M - 0.003467\frac{B}{T} + 0.3696C_{WP}\right] \tag{2.5.10}$$

The flat plate resistance is subsequently adjusted by including a form factor k during the calculation of total resistance. The constant c_{14} must be determined first to calculate form factor k, which serves the purpose of capturing the impact of the aft body shape.

$$c_{14} = 1.0 + 0.011C_{stern} \quad \textbf{with}$$

Aft body shape	C_{stern}
Pram with gondola	−25
V-shaped sections	−10
Normal sections	0
U-shaped sections	+10

(2.5.11)

Then, the form factor $(1 + k_1)$ can be determined with the constant c_{14}, the length of run L_R and input values from Table 2.2.

$$1 + k_1 = 0.93 + 0.487118c_{14}\left[\left(\frac{B}{L_{WL}}\right)^{1.06806}\left(\frac{T}{L_{WL}}\right)^{0.46106}\left(\frac{L_{WL}}{L_R}\right)^{0.121563}\left(\frac{L_{WL}}{V}\right)^{0.36486}(1 - C_p)^{-0.604247}\right] \tag{2.5.12}$$

Appendage Resistance

An appendage is defined as addition to the main part or main structure of a vessel (**Molland, 2011**). Examples of appendages include rudders, shaft brackets, skeg and bilge keels. The form factors associated with these appendages, denoted as k_{2_i} are presented in Table 2.3. In practice, reasonable estimates can be made based on these form factors, as model tests are not the most suitable method for accurately quantifying appendage resistance. Furthermore, effects of appendages are typically considered as a whole and not as individual unit (**Birk, 2019**).

The equivalent form factor for multiple appendages, $(1 + k_{2_i})_{eq}$ is given by:

$$(1 + k_{2_i})_{eq} = \frac{\sum_i (1 + k_{2_i}) S_{APP_i}}{\sum_i S_{APP_i}} \tag{2.5.13}$$

If bow thruster is present, the resistance due to the bow thruster tunnel R_{TH} can be obtained through:

Appendage	k_{2_i} value
rudder behind skeg	0.2 – 0.5
rudder behind stern	0.5
twin screw rudder (slender)	1.5
twin screw rudder (thick)	2.5
shaft brackets	2.0 – 4.0
skeg	2.0 – 3.0
strut bossing	2.0 – 3.0
hull bossing	1.0
exposed shafts (angle with buttocks about 10 degrees)	1.0
exposed shafts (angle with buttocks about 20 degrees)	4.0
stabiliser fins	1.8
dome	1.7
bilge keels	0.4

Table 2.3: Approximate values for appendage form factors k_{2_i}

$$R_{TH} = \rho v_S^2 \pi \mathrm{d}_{TH}^2 C_{D_{TH}} \quad \textbf{where} \quad C_{D_{TH}} = 0.003 + 0.003\left(\frac{10_{d_{TH}}}{t} - 1\right) \tag{2.5.14}$$

The coefficient $C_{D_{TH}}$ defines the drag coefficient for the tunnel, and it ranges between 0.003 and 0.012. Smaller values indicate thrusters which are in the cylindrical part of bulbous bow. The coefficient can also be estimated using the equation by **Hollenbach (1999)** in Equation (2.5.14).

With that, the appendage resistance R_{APP} can be calculated using:

$$R_{APP} = \frac{1}{2}\rho v_S^2 (1 + k_{2_i})_{eq} C_F \sum_i S_{APP_i} + \sum R_{TH} \tag{2.5.15}$$

Wave Resistance

The estimation of wave resistance R_W is dependent on Froude number Fr, and it is subdivided into three categories.[5]:

$$R_W(Fr) = \begin{cases} R_{W_a}(Fr) & \textbf{if} \quad Fr \leqslant 0.4 \\ \text{Interpolation} & \textbf{if} \quad 0.4 < Fr \leqslant 0.55 \\ R_{W_a}(Fr) & \textbf{if} \quad Fr > 0.5 \end{cases} \tag{2.5.16}$$

[5] Due to the length of the equations, only the case for $Fr \leqslant 0.4$ will be discussed in this thesis. The formulation for R_W for other Froude number range, can be found in the work by **Holtrop (1984)** and **Birk (2019)**

Chapter 3

Research Methodology

In this chapter the methodology used to develop random forest model will be discussed. The details of fusion between AIS data, ECMWF and CMEMS data source used for training the model will be presented in Section 3.1. Suitable methodology application during data pre-processing will be described in Section 3.2. The selection for appropriate, domain knowledge based, feature selection will be explained in Section 3.2.2. The selection of the most optimal model hyperparameter for different tree-based model will be explained in Section 3.3.2. Different performance metrics is used to validate the model's generalisation capability, The underlying principle of the metrics is elaborated in Section 3.3.1. Summary of methodology application in this study is summarised in Section 3.3.3 and visually represented in figure Figure 3.1.

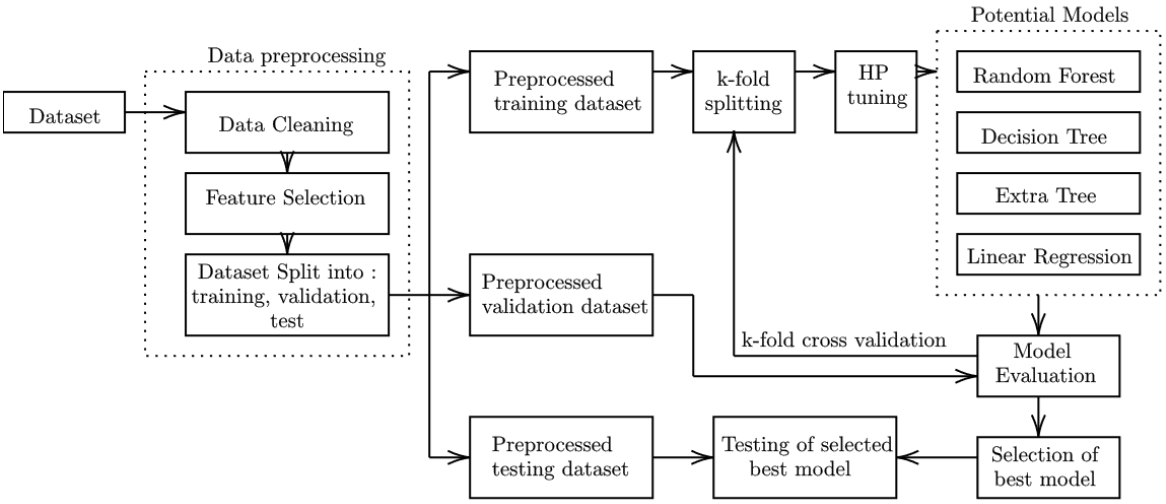

Figure 3.1: Scheme of proposed methodology

3.1 Data Acquisition

For the purpose of model training, 2021 AIS data from Ro-Ro ferry ship Hammershus is collected. The shore-based AIS data is made available by Danish Maritime

Authority which tracked her journey between ports of Køge, Rønne, Ystad and Sassnitz and structured according to Table 2.1. The AIS data is fused with weather data from ECMWF[1] with temporal resolution of 1 hour at granularity of 0.25° (longitude) x 0.25° (latitude), data from ECMWF provides information for wind, waves and seawater temperature. The information for current is obtained from CMEMS[2] with temporal resolution of 3 hours at granularity of 0.25° (longitude) x 0.25° (latitude).

The resulting fusion resulted in dataset with temporal resolution of 1 hour. Some information static information from the AIS data which only indicated the ship's identity are excluded. This includes ship's MMSI, Callsign, Name, IMO and Navigational Status. Additionally, information of the ship's Rate of Turn (ROT) is not available in this case. The weather information is synchronised so that the wind, waves, seawater temperature and sea current belongs to the same weather grid with same temporal resolutions.

The features (1) wind direction, (2) swell direction, (3) and wind wave direction are oriented to true north. However, to reflect the actual direction of weather effects that are acting on the ship, these features are converted to true direction; where true direction is defined as the direction of weather effect with respect to the bow of the ship. The value ranges between 0° and 180°. Subsequently, through vector decomposition, the northward and eastward wind velocity is converted to absolute wind speed and wind direction *with respect to True North*,φ:

$$V_{\text{wind}} = \sqrt{(V^N_{\text{wind}})^2 + (V^E_{\text{wind}})^2} \tag{3.1.1}$$

$$\varphi = \begin{cases} 360 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} > 0 \quad \wedge \quad V^N_{\text{wind}} < 0 \\ 180 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} < 0 \quad \wedge \quad V^N_{\text{wind}} > 0 \\ 270 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} > 0 \quad \wedge \quad V^N_{\text{wind}} > 0 \\ \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{otherwise} \end{cases} \tag{3.1.2}$$

Similarly, information of Northward and Eastward current Velocity is converted to absolute current speed and current direction *with respect to True North* γ.

$$V_{\text{current}} = \sqrt{(V^N_{\text{current}})^2 + (V^E_{\text{current}})^2} \tag{3.1.3}$$

[1] European Centre for Medium-Range Weather Forecast
[2] Copernicus Marine Environment Monitoring Service

Feature	Feature Name
AIS data	
Position Time Stamp [DD/MM/YYYY HH:MM:SS]	`Time`
Latitude [°]	`LAT`
Longitude [°]	`LON`
Width [m]	`width`
Length [m]	`length`
SOG [Knots]	`sog`
COG [m/s]	`cog`
Heading [°]	`heading`
Draught [m]	`draught`
Weather Data (0.5° Granularity)	
Wind Speed [m/s]	`windspeed`
True North Wind Direction, φ [°]	`truenorthcurrentdir`
Air Temperature Above Oceans [K]	`oceantemperature`
Maximum Wave Height [m]	`waveheight`
Swell Period [s]	`swellperiod`
Wind Wave Period [s]	`windwaveperiod`
Wave Period [s]	`waveperiod`
Sea Surface Temperature [K]	`surftemp`
Combined Wind Wave Swell Height [m]	`windwaveswellheight`
Swell Height [m]	`swellheight`
Wind Wave Height [m]	`windwaveheight`
Current Speed [m/s]	`curspeed`
True North Current Direction γ [°]	`truenorthcurrentdir`
True Wind Direction [°]	`truewinddir`
True Current Direction [°]	`truecurrentdir`
True Swell Direction [°]	`trueswelldir`
True Wind Wave Direction [°]	`truewindwavedir`
True Wave Direction [°]	`truewavedir`

Table 3.1: Structure of fused dataset

$$\gamma = \begin{cases} 360 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} < 0 \quad \wedge \quad V^N_{\text{current}} > 0 \\ 180 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} > 0 \quad \wedge \quad V^N_{\text{current}} < 0 \\ 270 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} < 0 \quad \wedge \quad V^N_{\text{current}} < 0 \\ \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{otherwise} \end{cases} \tag{3.1.4}$$

This conversion is performed as the information of current speed and current direction, γ, is necessary to perform the correction formula shown in Equation (2.3.5) and Equation (2.3.6). However, for training purpose, this feature will not be considered. Instead, the true current direction and true wind direction will be considered. The initial structure have 27 features, 9 AIS features and 18 weather features. The structure of the initial dataset i.e. before data preprocessing and feature selection, is summarised in Table 3.1

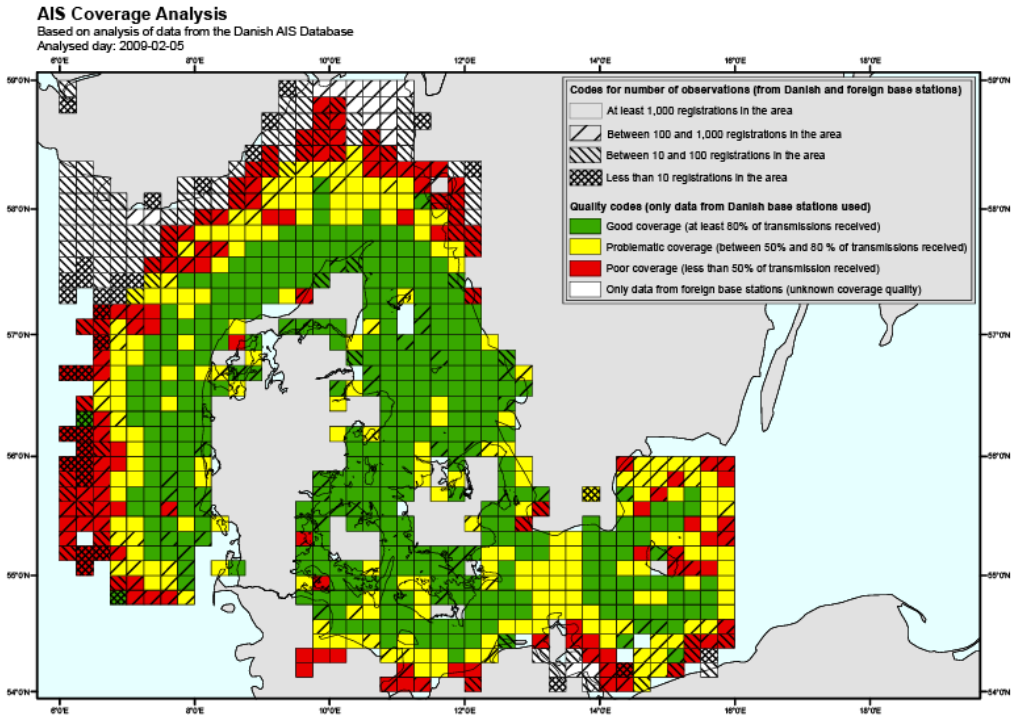

Figure 3.2: Shore based AIS Coverage based on data from AIS database Danish Maritime Authority (2023)

3.2 Data Preprocessing

This section presents the steps taken to during data preprocessing. The dataset will be first subjected to data cleaning which include identification of anomalies and missing values, the steps are explained in Section 3.2.1. Boundary condition is then applied to ensure that the model represent operating condition at steady state. Using domain knowledge, appropriate features are selected and discarded to ensure the model obeys shipping domain knowledge. This dataset is to be split into training, validation and test dataset. These steps will be further elaborated in Section 3.2.2.

3.2.1 Data Cleaning

The journey between the port of Køge, Rønne, Ystad and Sassnitz is plotted using `QGIS`[3]. The plot of the journey is shown in Figure 3.3, it can be seen, that the journey between Rønne and Sassnitz is not represented completely. As in this information is missing due to poor coverage in the area between Sassnitz and Rønne. This is shown by the plot shown in Figure 3.2. Therefore, the data plot for the journey between Sassnitz and Rønne will be excluded. Basic threshold of decimal degrees of 55.04° N for latitude is applied, this threshold will exclude the journey between Sassnitz and Rønne.

In its initial state, the dataset contains 7453 data points which described the journey

[3] https://qgis.org/en/site/, `QGIS` is a free and open source geographic information system

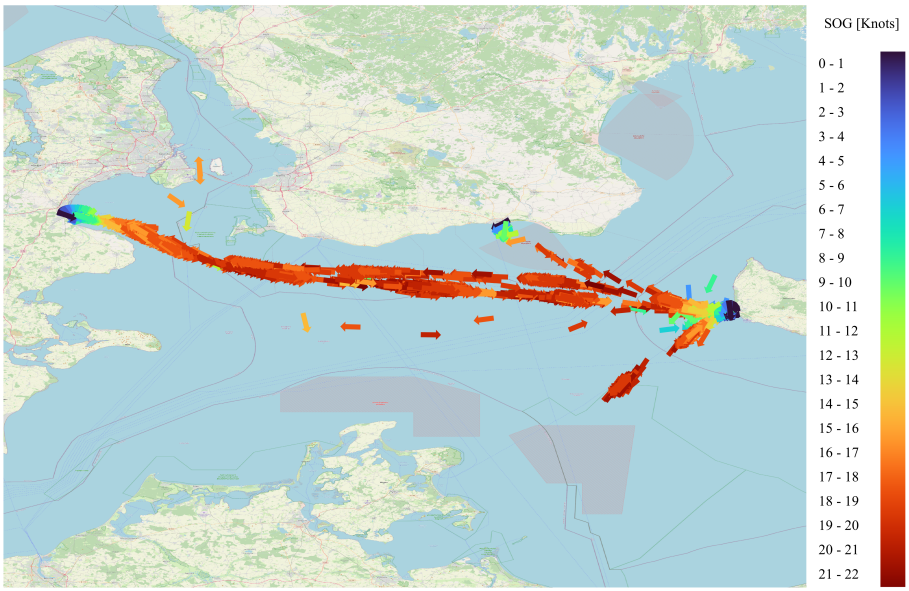

Figure 3.3: Journey of the ship in a year

of the ship in one year. The initial data points represented all navigational status of the ship, which include "mooring", "anchoring" and "underway using engine". This is clearly observed in the histogram for the SOG distribution in figure BLALA. To ensure that the dataset represents the actual operating condition of ship in steady state, a threshold of 5 knots is applied. SOG can vary due to changing sea state, but it can also be reduced by the ship's operator around the port when it departs from port of origin or arriving at port of arrival. Any data points with SOG less than 5 knots will be discarded which is considered as manoeuvring Abebe et al. (2020). After applying the SOG threshold, the amount of data points significantly decrease from 7453 data points to 3506 data points. This indicated that about half of the total data points represented the ship's stationary behaviour.

From preliminary analysis, possible source of error is identified for data points representing current speed. In range of current speed between 0.01 and 0.03 [m/s], noticeable peak in data points is observed. This peak attributed to missing information on northward and eastward current speed in some data points from the provided dataset. This resulted in single random error value for current speed which resulted in the peak observed in the histogram.

To address the missing values, the missing values for eastward current and northward current are imputed using `KNNImputer` feature from `Scikit-Learn`. Each sample's missing values are imputed using the mean of nearest neighbour found in training dataset Fabian Pedregosa et al. (2011). Once the missing values of northward and southward current are imputed, the current speed for the missing values will be recalculated.

The imputing approach using k-nearest neighbour is also applied to other weather features that contained missing values i.e. `NaN` values. Imputing missing values

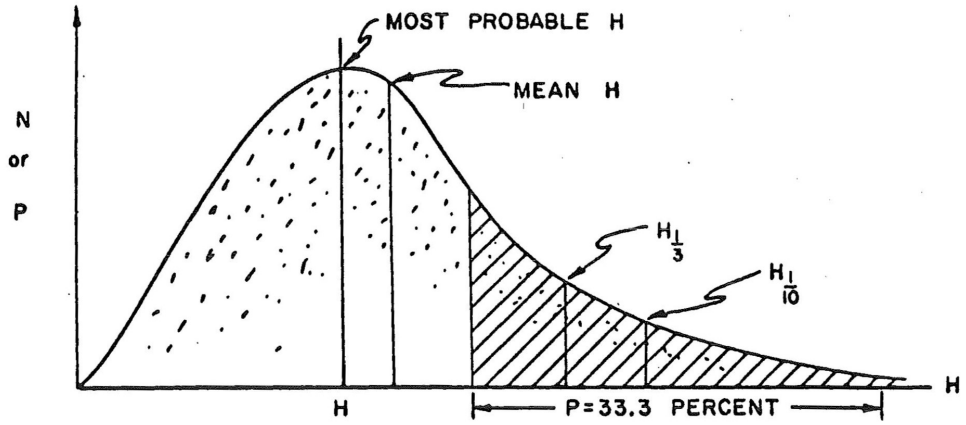

Figure 3.4: Statistical distribution of wave heights Bretschneider (1965)

is necessary as modelling package by `Scikit-Learn` cannot handle missing values. Imputing strategy using k-nearest neighbour is considered as it should reflect the weather conditions within the region of missing values.

3.2.2 Feature Selection

To select appropriate features for the model, correlation between the features is first studied. Feature selection is necessary to simplify the model and subsequently save computing cost during training. Selection of features is based on statistical approach of High Correlation Filter proposed by Abebe et al. Abebe et al. (2020). This approach considers pairs of features with correlation features higher than 0.7 as one entity. However, the selection of highly correlated features must not violate natural state of matter. Therefore, in addition to statistical approach, the scientific reasoning behind the correlations will be considered and prioritised over the statistical approach.

From AIS data, the information on (1) time, (2) latitude, (3) longitude, (4) width, and (5) length are not included for training. As time, latitude and longitude have no impact on the ship. While the width and length is properties from the ship that remain constant.

The features (1) combined wind wave swell height, (2) swell height, maximum wave height (3) and wind wave height are physically correlated. In sea wave theory, wind wave swell height is also known as significant wave height $H_{1/3}$. It is defined as the mean of the highest one-third of waves in the wave record Holthuijsen (2007).

The distribution of wave heights can be represented by probability density function. Hence, the term "highest one-third of waves" here means the region of wave heights that belong in the upper one-third of a probability density function, this is illustrated in Figure 3.4. From this distribution, the relation between significant wave height $H_{1/3}$, the highest ten percent of waves H_{10} and average wave height $\overline{H}$ can

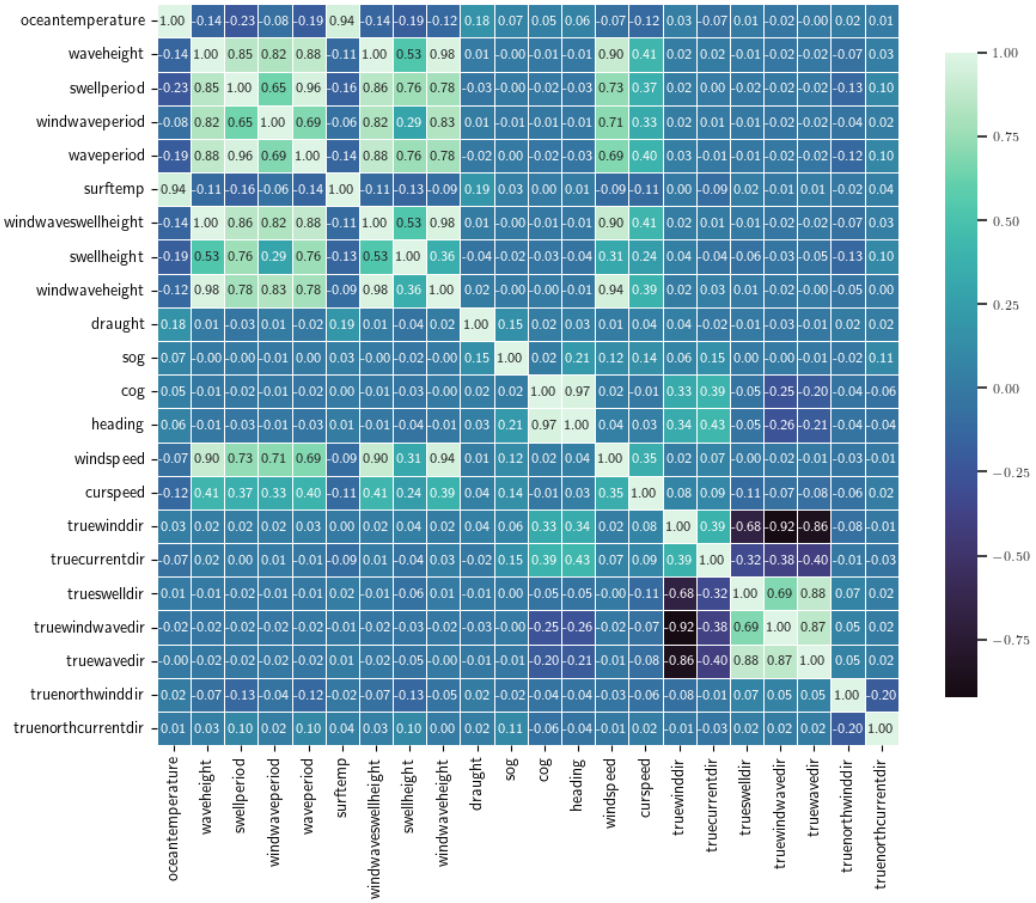

Figure 3.5: Correlation Heat Map

be summarised as follows Bretschneider (1965); Holthuijsen (2007):

$$\overline{H} = 0.625 \cdot H_{1/3} \tag{3.2.1}$$

$$H_{10} = 2.03 \cdot \overline{H} = 1.27 \cdot H_{1/3} \tag{3.2.2}$$

$$H_{\max} = 2 \cdot H_{1/3} \tag{3.2.3}$$

Additionally, Bitner-Gregersen Bitner-Gregersen (2005) described the relation between the significant wave height, wind wave height and swell height through following equation:

$$H_{1/3} = \sqrt{(H_{\text{swell}})^2 + (H_{\text{windwave}})^2} \tag{3.2.4}$$

From here, it is clear that significant wave height should be retained for modelling, as it holds critical information regarding wave properties. The features swell height, wind wave height and maximum wave height will be dropped as it can be defined through correlations defined in Equation (3.2.1),Equation (3.2.2),Equation (3.2.3) and Equation (3.2.4). This decision is also statistically supported through the high correlation filter method. As shown in Figure 3.13, high correlation are observed between these features.

Training Label	
SOG [Knots]	`sog`
Training Features	
COG [m/s]	`cog`
Heading [°]	`heading`
Draught [m]	`draught`
Wind Speed [m/s]	`windspeed`
Air Temperature Above Oceans [K]	`oceantemperature`
Maximum Wave Height [m]	`waveheight`
Wave Period [s]	`waveperiod`
Sea Surface Temperature [K]	`surftemp`
Combined Wind Wave Swell Height [m]	`windwaveswellheight`
Current Speed [m/s]	`curspeed`
True Wind Direction [°]	`truewinddir`
True Current Direction [°]	`truecurrentdir`
True Wave Direction [°]	`truewavedir`

Table 3.2: Structure of fused dataset

From Figure 3.13, high correlation is observed between wave period, swell period and wind wave period. Bitner-Gregersen further elaborated that the state of the sea can be described through the significant height $H_{1/3}$ and spectral peak T_p with help of Torsethaugen peak K. Torsethaugen et al. (2004). Hence, the features swell period and wind wave period are discarded as it only distinguish whether the sea is dominated by swell or by wind. The feature wave period will still be retained. Consequently, the features true wind wave direction and true swell direction will be discarded as the features that explained the magnitude of these features are discarded.

Statistically, the heading and COG are highly correlated, but both features are retained as it explain two different parameters of the ship. Course Over Ground reflects the ship course heading while heading represented the actual heading of the ship at a particular point of time. Same principle also apply between air temperature above ocean and sea surface temperature. Air temperature above oceans represents the temperature of wind while sea surface temperature represents current temperature of current.

From feature selection, 5 features from AIS data are discarded while 11 features are removed from the weather data. To predict the ship speed, The SOG will be selected as the label to train the model. The remaining attributes will be selected as training features. This is summarised in Table 3.2.

3.3 Modelling

In this section, the modelling of ship speed through SOG using selected features will be performed using tree-based regressor model. The tree-based regressor model considered are decision tree regressor, random forest regressor and extra-tree regressor. In addition, the tree-based models are compared against multiple linear regressor to as benchmark. The methodology to develop the best model is divided into several steps.

For training, the dataset is split into training, validation and test dataset in ratio of 73:18:9. Journey data from the month of June is arbitrarily selected as test dataset. The remaining dataset will be split into training and validation dataset in 80:20 ratio. The explanation of training process and selection of the best model is broken down into several sections. In Section 3.3.2, the tuning parameter of `Scikit-Learn` will be studied extensively as suitable tuning could result in improved model performance.

Appropriate statistical performance measures are applied to each model; the performance measures selected will help to evaluate how well a model is able to make generalisation on validation and test dataset. The evaluation will be cross validated in form of k-folding. The details on evaluation methodology used in this thesis will be discussed in Section 3.3.1.

3.3.1 Performance Metrics for Validation

To gain sensible estimate of model performance and how precise a model is, the model will be cross validated by means of k-folding. K-fold cross validation split the training set into k subsets which is called *folds*, then the model will be trained k times using k-1 subsets and remaining one for validation, this process is illustrated in Figure 3.6. For each iteration, each model is evaluated using different performance metrics such as (1) Coefficient of Determination (R^2), (2) Explained Variance (EV), (3) Mean Absolute Error (MAE), (4) Root Mean Square (RMSE) and (5) Median Absolute Deviation (MAD). The results from each iteration is then averaged, where the information on model precision can be gained from the standard deviation. Performing k-fold cross validation checks model robustness against different datasets. The properties of each performance metric will be discussed in the following sections.

Coefficient of Determination (R^2)

The coefficient of determination R^2 gives a measure on prediction quality, R^2 quantifies the ability of the regression model to approximate the actual values. R^2 is defined by Equation (3.3.1), where y represents true target output, $\hat{y}$ represents the predictor output and $\overline{y}$ represents the mean. R^2 score range between 0 and 1, higher values i.e. $R^2 \rightarrow 1$ indicate better model fit and score of 1 indicate perfect prediction.

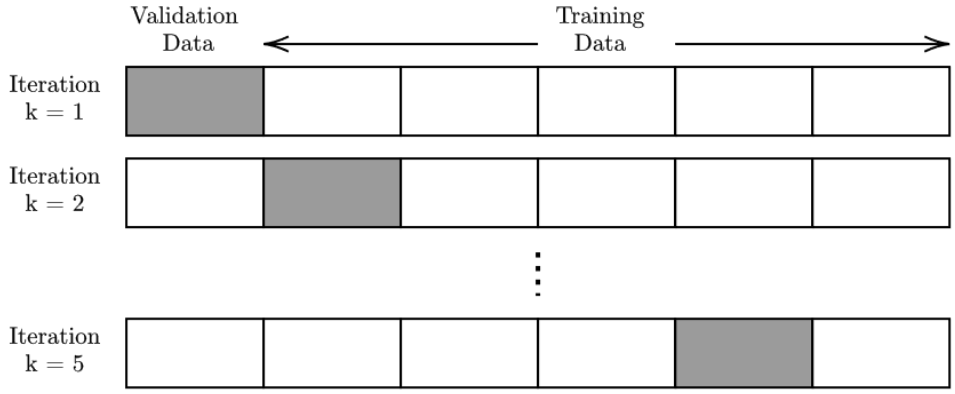

Figure 3.6: Visual illustration of k-folding, Grey shaded box represents the validation data while white box represents the training data

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_\mathrm{i} - \hat{y}_\mathrm{i})^2}{\sum_{i=1}^{n}(y_\mathrm{i} - \overline{y}_\mathrm{i})^2} \quad \textbf{where} \quad \overline{y} = \frac{1}{n}\sum_{1}^{n} y_\mathrm{i} \tag{3.3.1}$$

Explained Variance (EV)

Explained variance indicate how well a model can capture variance from a dataset. It is defined by Equation (3.3.2), where σ_x represents standard deviation of parameter x. EV score range between 0 and 1, where the best score of $EV = 1$ can be obtained if $\sigma^2_{(y-\hat{y})} \to 0$.

$$EV(y,\hat{y}) = 1 - \frac{\sigma^2_{(y-\hat{y})}}{\sigma^2_y} \tag{3.3.2}$$

Mean Absolute Error (MAE)

MAE indicated the expected value of absolute (L^1 norm) error, and it can be calculated by:

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_\mathrm{i} - \hat{y}_\mathrm{i}| \tag{3.3.3}$$

Root Mean Square Error (RMSE)

The RMSE describe the expected value of quadratic error. RMSE place large penalty on large deviation between true and estimated values and for this reason, it can be used to as a metric to indicate model performance against outliers. Ideal score is observed when RMSE $\to 0$. RMSE can be considered as absolute measure of model fitness. Omitting the root term, RMSE becomes MSE, which is the loss function of

Equation (2.2.2) that is used to determine the most optimal split in a regression decision tree.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3.4}$$

Median Absolute Deviation (MAD)

MAD is a performance metrics that considers the median of the absolute errors. It is robust to outlier as it only consider median performance

$$MAD(y, \hat{y}) = \text{median}(|y_1 - \hat{y}_1|, \ldots, |y_i - \hat{y}_i|) \tag{3.3.5}$$

3.3.2 Model Hyperparameter Optimisation

The subject of parameter tuning was briefly discussed in Section 2.2.1. In Section 2.2.1 parameter tuning was applied to decision tree regressor to avoid overfitting by changing the minimum amount of samples a leaf node has. This example implies that altering model hyperparameter will affect the model performance. However, the optimisation of the hyperparameter cannot be performed *a priori* and as such iterative process will be performed until best hyperparameter value is found.

Model	Decision Tree	Random Forest	Extra-Trees
Number of trees	1	Many	Many
Features considered for split at each node	All features	Random subset of features	Random subset of features
Bootstrapping	Not applied	Yes	No
Split Rule	Best split	Best split	Random split

Table 3.3: Comparison of tree based model from Section 2.2

`Scikit-Learn` offers `GridSearchCV` and `RandomizedSearchCV` to help search for the most optimal hyperparameter. Both solutions operate with similar principle: The selected hyperparameters to be tuned with its value range is evaluated using cross validation to evaluate the best possible combination between the selected hyperparameters. The difference between `GridSearchCV` and `RandomizedSearchCV` lies in how it searches for the best value for the selected hyperparameters: `GridSearchCV` involves construction of grids containing all possible combinations of hyperparameter value in specified range.`RandomizedSearchCV` randomly samples hyperparameter values.

The exhaustive nature of `GridSearchCV` means that it is computationally costly to perform, especially when there are multiple hyperparameters to be considered and value search space is large. `RandomizedSearchCV` gives more control to computing budget by setting the number of iteration and usually produces more accurate results than `GridSearchCV` approach. Géron (2019); Bergstra and Bengio (2012).

For this reason, the `RandomizedSearchCV` will be employed to search for best possible hyperparameter. However, the limitation of *a priori* knowledge of hyperparameter value still exists. In spite of `RandomizedSearchCV` ability to control the computational budget, it is still takes considerable time to obtain the best hyperparameter value. The computational budget may be spent on searches in unpromising search space. With that, initial exploration on the effect of each hyperparameter on model performance will be performed to give better overview on which search space that should be considered during hyperparameter optimisation. In the next subsections, the effect of tunable hyperparameter of tree-based model from `Scikit-Learn` will be explored to give baseline numbers for the search space. RMSE is used as performance metrics as the hyperparameter parameter optimisation done in this thesis aims to reduce the error during prediction.

3.3.2.1 Number of features

Defined with default value as `max_features=None` in `Scikit-Learn`. This hyperparameter controls the number of features to be considered when looking for the best split, the default `None` option means it will consider all features. This parameter tuning is available for Decision Tree Regressor, Random Forest Regressor and Extra-Tree Regressor. Initial exploration indicated Random Forest Regressor and Extra Tree Regressor benefit from considering more features, Decision Tree Regressor requires further fine-tuning to optimise the model as the default `None` means it will consider all features when searching for best split.

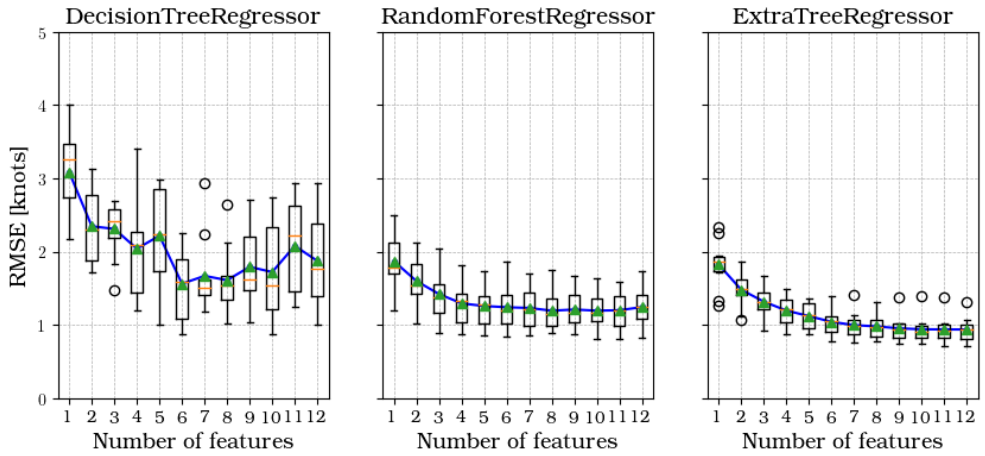

Figure 3.7: Hyperparameter tuning of `max_features`

3.3.2.2 Number of sample in a leaf node

Defined with default value as `min_samples_leaf=1` in `Scikit-Learn`. This parameter controls number of samples required to be at leaf node, where split point will be considered if the leaf contains at least `min_samples_leaf=n` training samples in each left and right branch. As shown in Figure 2.4, tuning this hyperparameter to higher

values helps to smoothen the model and avoid overfitting. However, this may lead to underfitting as the model is unable to capture the trend within the data. This is supported by the findings shown in Figure 3.8, the DTR benefits from regularisation at certain breakeven point, in this case, it is found to be at `min_samples_leaf=4`. But after this breakeven point, the model's performance degrades. It is also observed that RFR and ETR does not benefit from any form of regularisation.

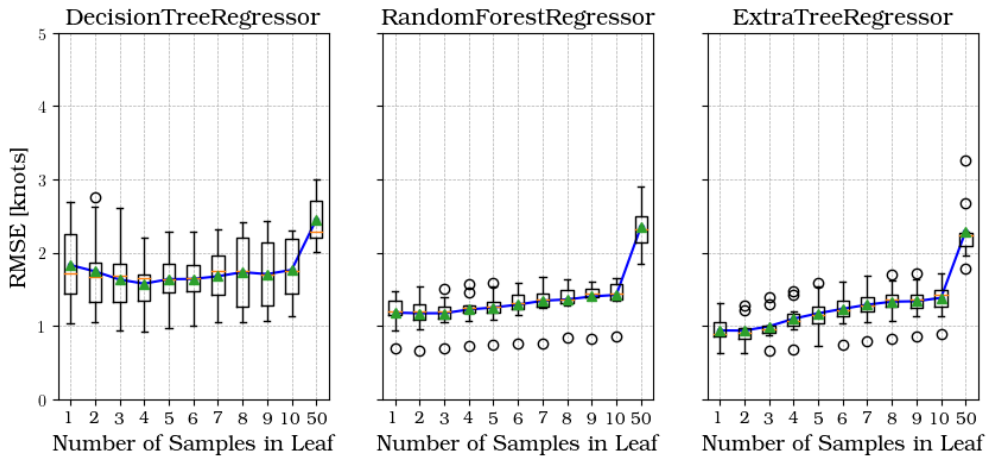

Figure 3.8: Hyperparameter tuning of `min_samples_leaf`

3.3.2.3 Depth of Tree

Defined with default value as `max_depth=None` in `Scikit-Learn`. This hyperparameter controls the growth of the tree. Leaving it at `max_depth=None` means the tree will grow until all leaves are pure i.e. until minimum MSE is obtained or when the number of samples is less than the minimum number of samples required to split an internal node. Similar to `min_samples_leaf`, DTR shows improvement until a certain breakeven point. RFR performance seems to stabilise at certain depth while ETR benefits from allowing full growth of the tree. The results are summarised in Figure 3.9

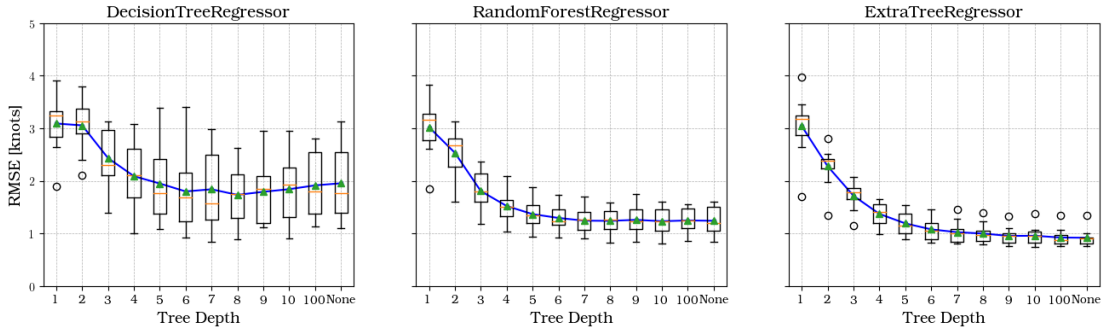

Figure 3.9: Hyperparameter tuning of `max_depth`

3.3.2.4 Number of Trees

Defined with default value as `n_estimators=100`. This hyperparameter controls the amount of trees i.e. predictors in a forest. Tuning of number of trees will have an effect on the training time and it is only available to RFR and ETR. The default value seems to yield satisfactory result, as the performance for both RFR and ETR stabilise after in this case stabilise after 100 trees, as seen in Figure 3.7.

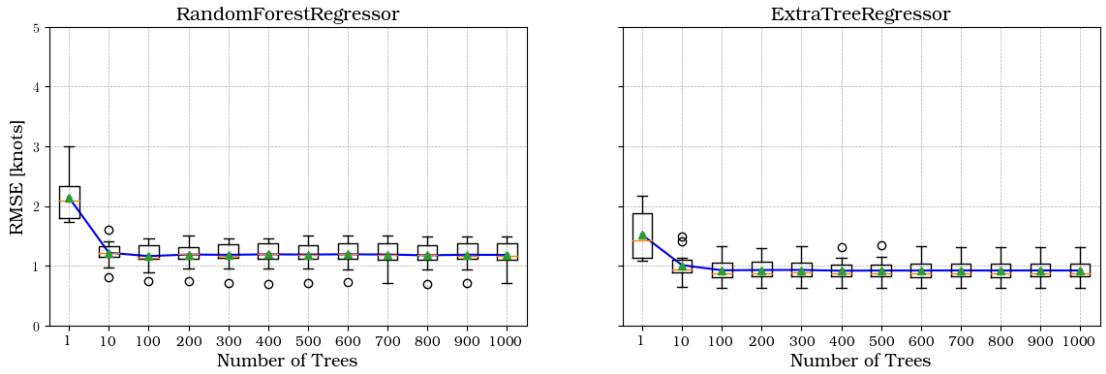

Figure 3.10: Hyperparameter tuning of `n_estimators`

3.3.3 Methodology Application

As such, this thesis aims to find optimisation possibilities for the BBM to extract maximum prediction performance from tree-based model. The estimation of engine power using Holtrop-Mennen method involves a lot of The approximations. To ensure correctness during estimation of engine power, the approximations are based on of ship dimensions and mechanical data are based

- Two data sources are imported. `AIS_weather_H_ok2_copy.csv` and `AIS_weather_h_rename_copy.csv`. The information from the latter comma delimited file will be used for calculating the ship Speed Through Water (STW). The information required is the true north current direction. Which is obtained from the vector component of the Northward and Southward current.
- This dataframe will be merged with the main dataframe from the file `AIS_weather_H_ok2_copy.csv`.
- Omission of the journey data between Ronne and Sassnitz
- SOG threshold is applied to omit ship mooring and maneuvering to accurately represent the ship's steady state operation Abebe et al. (2020); Bal Beşikçi et al. (2016); Gkerekos et al. (2019); Yang et al. (2020). This threshold is selected as 5 knots according to Abebe et al. (2020)
- The AIS data from June is filtered. This data will be used as validation data to check the model's performance.

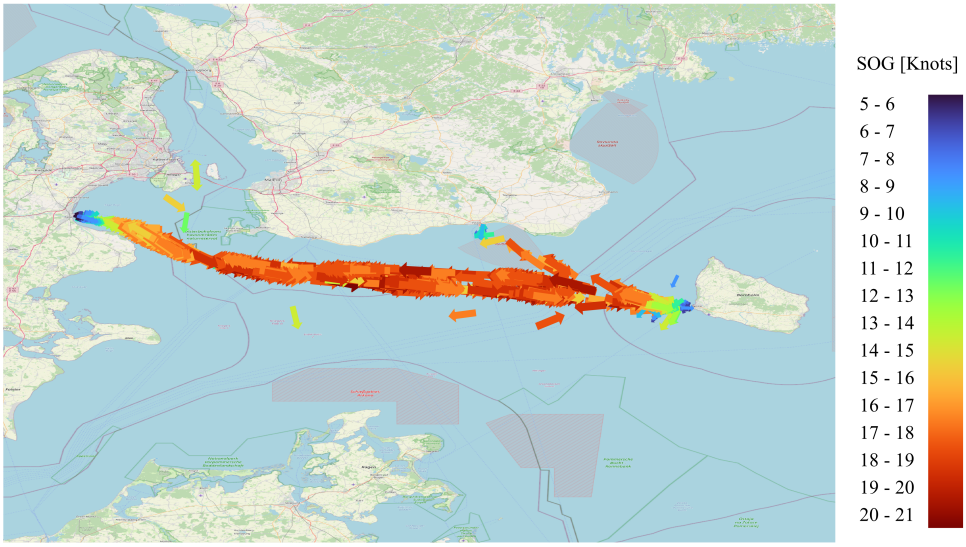

Figure 3.11: Journey of the ship in June

3.3.4 Data Analysis

- The features are represented in a histogram plot. For the feature Current speed, anomaly is detected. Certain spike is detected around $0.01 - 0.03$ `m/s`. Reasons unknown. The data is retained, including the spike, until a definitive answer can be found.
- OPEN QUESTION : What is the necessity of feature standardization / normalization ? Normalization is required for ANN as model training requires the value between 0 and 1. But in case of RFR, there is no such requirement. Through testing, data standardization also does not seem to improve the model's performance.

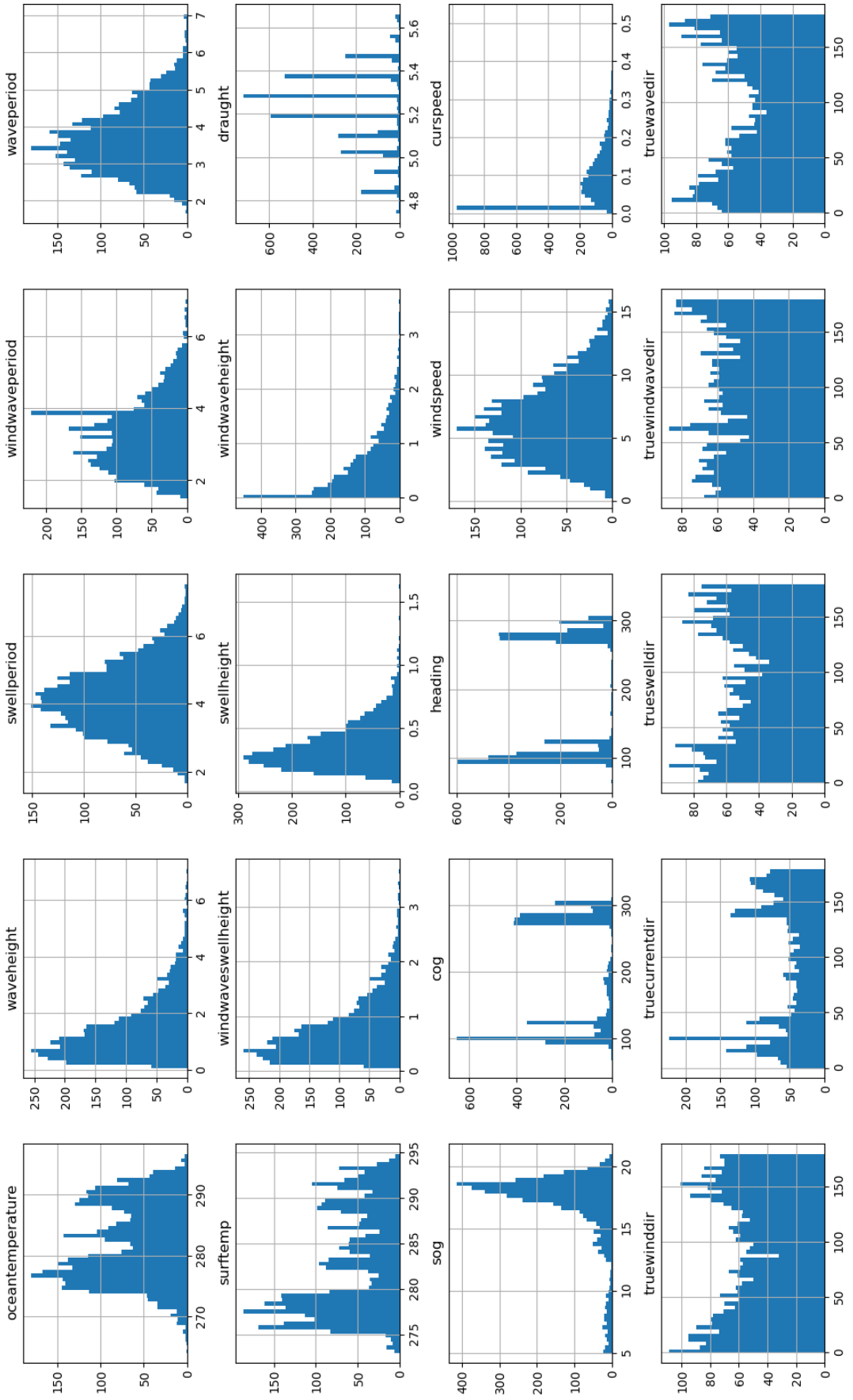

Figure 3.12: Histogram of the features

- The correlation of the features against SOG are determined. It is found that :
  - Draught
  - Course Over Ground (COG)
  - heading
  - Wind Speed
  - Current Speed
  - True Current direction

  Have relatively stronger correlation to SOG compared to other features, albeit the correlation is a weak one
- The correlation between the features is displayed using the following the heat map. From the heat map it can be observed that between these features:
  - Waveheight and wind wave swell height
  - Waveheight and wind wave height
  - Windwaveswellheight and wave period

  Have a strong correlation between each other.
- Open topic:
  - Feature reduction is possible, Abebe et al. (2020) suggested high feature correlation filter, the filter suggest that two features which has a high correlation (> 90%) is to be combined into a single feature. But the author is unsure whether this combination is physically sensible. Hence, this filter is yet to be applied for feature reduction.
  - Some of these features can be connected through wave equations, but the author has not found an equation which could relate these features.
- The random forest regressor could not function when `NaN` values are present. With that, the missing values are filled in using the `imputer` function. The missing values are filled in by means of `KNN`.

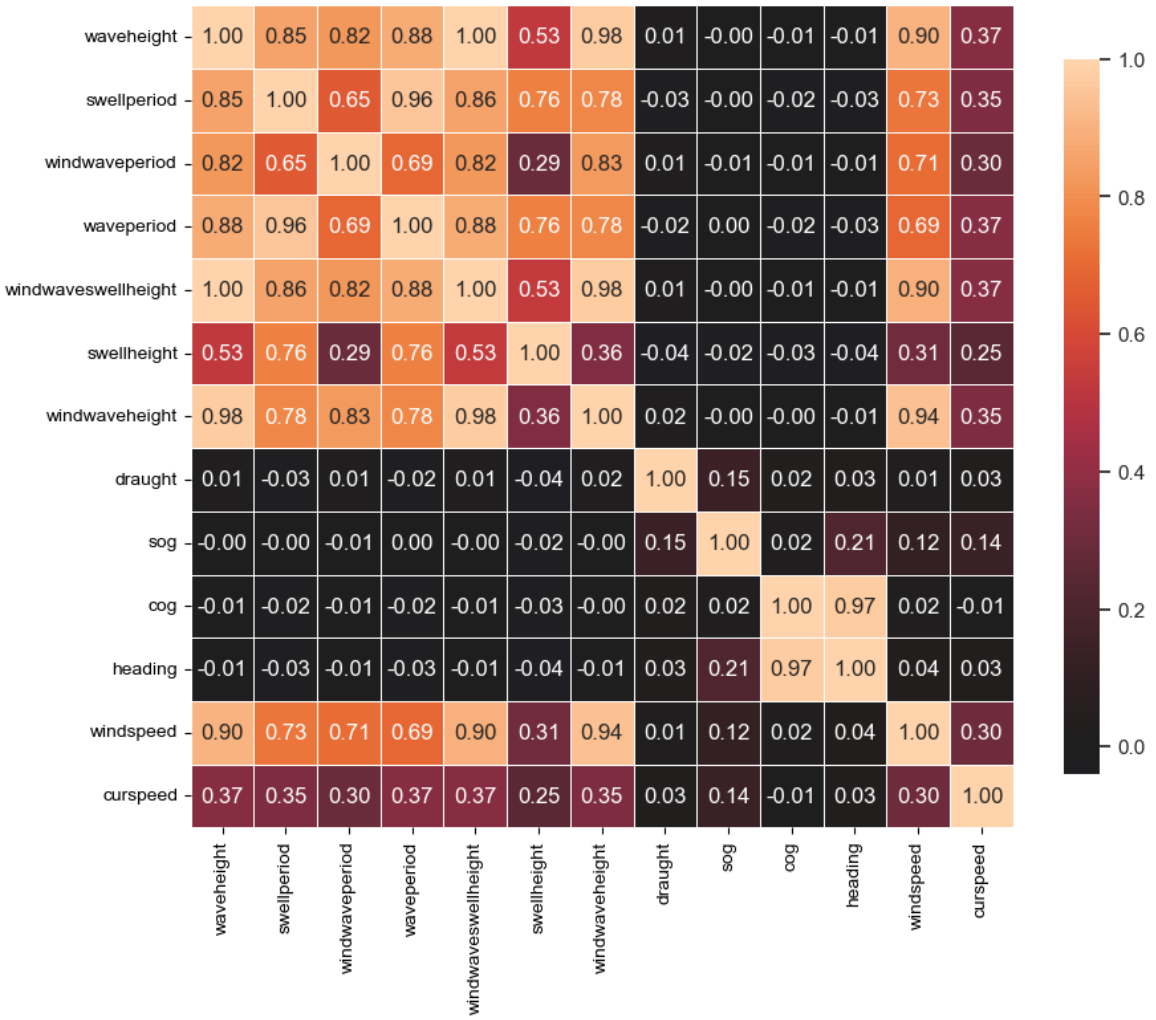

Figure 3.13: Correlation Heat Map

3.3.5 Modelling

- The data is split into 80:20 ratio. But considering the validation data, it is split into approximately 73:18:9.
- The model is then trained using Random Forest Regression (RFR). Additional training is also performed using Decision Tree Regressor (DTR). DTR model performance will be used as a benchmark as it is also a tree-based modelling method with similar methodology to RFR.
- The computational time of DTR is significantly faster than RFR Model Evaluation

3.3.6 Predicting STW

- The ship's Speed Through Water STW can be calculated using vector component of the SOG and current speed. The direction used will be according to True North. Yang et al. (2020); Zhou et al. (2020)

- SOG represents the speed of the ship with reference to the ground, while the STW represent the ship's speed with reference to water.
- SOG also can be termed by the ship's speed that is captured by the GPS, and does not consider any effect of the current
- This means that the ship's STW will be greater than the ship's SOG when there is current moving against the ship's movement direction and vice versa
- The vector decomposition can be defined from the following equations, which is based on the equation by Yang et al. (2020):
  - The ship's SOG V_g can be decomposed into V_g^x and V_g^y, which represents the x and y components of the SOG respectively using the ship's course heading (COG) β *with respect to True North*:

$$V_g^x = V_g \sin(\beta) \tag{3.3.6}$$

$$V_g^y = V_g \cos(\beta) \tag{3.3.7}$$

  - To consider the effect of sea current. The current speed V_c will also be decomposed to x and y components respectively using the current direction γ *with respect to True North*:

$$V_c^x = V_g \sin(\gamma) \tag{3.3.8}$$

$$V_c^y = V_g \cos(\gamma) \tag{3.3.9}$$

  - from here the ship' STW V_{wx} and V_{wy} component can be found from the following equation:

$$V_w^x = V_g^x - V_c^x \tag{3.3.10}$$

$$V_w^y = V_g^y - V_c^y \tag{3.3.11}$$

  - The magnitude of the STW can be readily obtained from the following vector synthesis

$$V_w = \sqrt{(V_w^x)^2 + (V_w^y)^2} \tag{3.3.12}$$

- This principle is applied to the following Python script. 3.3.8

```
1
2       # Convert SOG from [Knots] to [m/s]
3
4       dfprog["vgms"] = dfprog["sog_pred"]/1.9438
5
6       # Convert the angles from [Degrees] to [Radians]
7
8       rad_gamma = np.deg2rad(dfprog["gamma"])
9       rad_cog = np.deg2rad(dfprog["cog"])
10
11      # Decomposition in x-component
12
13      dfprog["vgx"] = dfprog["vgms"] * np.sin(rad_cog)
14      dfprog["vcx"] = dfprog["curspeed"] * np.sin(rad_gamma)
15      dfprog["stw_x"] = (dfprog["vgx"] - dfprog["vcx"])
16
17      # Decomposition in y-component
18
19      dfprog["vgy"] = dfprog["vgms"] * np.cos(rad_cog)
20      dfprog["vcy"] = dfprog["curspeed"] * np.cos(rad_gamma)
21      dfprog["stw_y"] = (dfprog["vgy"] - dfprog["vcy"])
22
23      # Vector synthesis and reconversion to [Knots] from [m/s]
24
25      dfprog["vwms_p"] = np.sqrt(dfprog["stw_x"]**2 + dfprog["stw_y"]**2)
26      dfprog["stw_pred"] = dfprog["vwms_p"]*1.9438
27
28
```

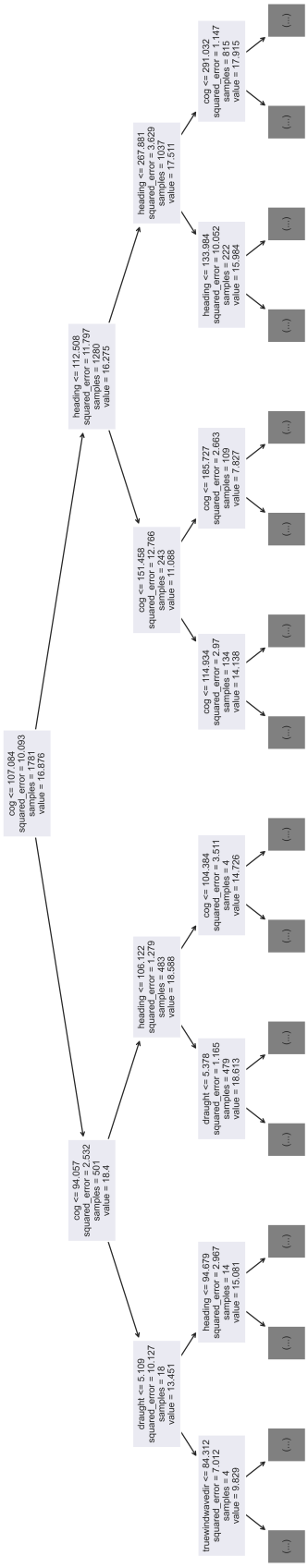

Figure 3.14: Correlation Heat Map

Chapter 4

Result and Discussion

The result of the research is discussed in this chapter. This comprises model validation and how different statistical metrics are used to analyze the model's performance.

4.0.1 Model Evaluation

The model are tested against four metrics, namely:

- R^2 : Indicate model fit. Best Score = 1
- Explained Variance `EV` : Indicate amount of variance in model. Best Score = 1
- Mean Absolute Error `MAE` : Indicate how much error a model makes in its prediction. Best Score = 0
- Root Mean Square Error `RMSE` : Same as MAE, more sensitive to outlier. Best Score = 0
- Median Absolute Error `MAD` : Check robustness against outlier. Best Score = 1

The result is summarized in the following table

Model	RFR	DTR	LR
R^2	0.9328181446941499	0.8526085810220092	1
EV	0.932872958708872	0.8526260247615258	2
MAE	0.5546347329650284	0.8108982427834758	3
RMSE	0.7095480848510665	1.5566896535262504	4
MAD	0.38484635910000087	0.5475717149999983	5

Table 4.1: Model performance

Model	RFR	DTR	LR
R^2	0.9328181446941499	0.8526085810220092	1
EV	0.932872958708872	0.8526260247615258	2
MAE	0.5546347329650284	0.8108982427834758	3
RMSE	0.7095480848510665	1.5566896535262504	4
MAD	0.38484635910000087	0.5475717149999983	5

Table 4.2: Model performance

Chapter 5

Summary and Outlook

In this chapter the summary of this research will be discussed. This section includes reflections of the research process and presents any possible suggestions and recommendations in this line of research. This chapter concludes this thesis.

Bibliography

Misganaw Abebe, Yongwoo Shin, Yoojeong Noh, Sangbong Lee, and Inwon Lee. Machine learning approaches for ship speed prediction towards energy efficient shipping. *Applied Sciences*, 10(7):2325, 2020. doi:10.3390/app10072325.

E. Bal Beşikçi, O. Arslan, O. Turan, and A. I. Ölçer. An artificial neural network based decision support system for energy efficient ship operations. *Computers & Operations Research*, 66:393–401, 2016. ISSN 03050548. doi:10.1016/j.cor.2015.04.004.

J Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 2012. URL https://www.semanticscholar.org/paper/Random-Search-for-Hyper-Parameter-Optimization-Bergstra-Bengio/188e247506ad992b8bc62d6c74789e89891a984f.

Nicolas Bialystocki and Dimitris Konovessis. On the estimation of ship's fuel consumption and speed curve: A statistical approach. *Journal of Ocean Engineering and Science*, 1(2):157–166, 2016. ISSN 24680133. doi:10.1016/j.joes.2016.02.001.

Lothar Birk. *Fundamentals of ship hydrodynamics: Fluid mechanics, ship resistance and propulsion / Lothar Birk*. John Wiley & Sons, Hoboken, New Jersey, 1st edition, 2019. ISBN 1118855515. doi:10.1002/9781119191575. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119191575.

Elzbieta M. Bitner-Gregersen. Joint probabilistic description for combined seas. In *24th International Conference on Offshore Mechanics and Arctic Engineering: Volume 2*, pages 169–180. ASMEDC, 2005. ISBN 0-7918-4196-0. doi:10.1115/OMAE2005-67382.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi:10.1023/A:1010933404324. URL https://link.springer.com/article/10.1023/a:1010933404324.

Charles L. Bretschneider. *Generation of waves by wind. State of the art*. 1965. URL https://apps.dtic.mil/sti/citations/ad0612006.

Andrea Coraddu, Luca Oneto, Francesco Baldi, and Davide Anguita. Vessels fuel consumption forecast and trim optimisation: A data analytics perspective. *Ocean Engineering*, 130:351–370, 2017. ISSN 00298018.

doi:10.1016/j.oceaneng.2016.11.058. URL https://www.sciencedirect.com/science/article/pii/S0029801816305571.

Danish Maritime Authority. Safety at sea, navigational information, AIS data, 2023. URL https://dma.dk/safety-at-sea/navigational-information/ais-data.

Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000. ISSN 1573-0565. doi:10.1023/A:1007607513941. URL https://link.springer.com/article/10.1023/A:1007607513941.

Yuquan Du, Qiang Meng, Shuaian Wang, and Haibo Kuang. Two-phase optimal solutions for ship speed and trim optimization over a voyage using voyage report data. *Transportation Research Part B: Methodological*, 122:88–114, 2019. ISSN 01912615. doi:10.1016/j.trb.2019.02.004.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems / Aurélien Géron*. O'Reilly, Sebastopol, CA, second edition edition, 2019. ISBN 978-1-492-03264-9.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. ISSN 1573-0565. doi:10.1007/s10994-006-6226-1. URL https://link.springer.com/article/10.1007/s10994-006-6226-1.

Christos Gkerekos, Iraklis Lazakis, and Gerasimos Theotokatos. Machine learning models for predicting ship main engine fuel oil consumption: A comparative study. *Ocean Engineering*, 188:106282, 2019. ISSN 00298018. doi:10.1016/j.oceaneng.2019.106282.

H. E. Guldhammer and S. A. Harvald. Ship resistance - effect of form and principal dimensions. (revised). *Danish Technical Press, Danmark, Danmarks Tekniske Hojskole, kademisk Forlag, St. kannikestrade 8, DK 1169 Copenhagen*, 1974. URL https://repository.tudelft.nl/islandora/object/uuid:4a6f2694-a3ab-4a90-beac-7f38c41d4e40.

Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. ISSN 1941-1294. doi:10.1109/MIS.2009.36.

Michael Haranen, Pekka Pakkanen, Risto Kariranta, and Jouni Salo. White, grey and black-box modelling in ship performance evaluation. 2016. URL https://www.researchgate.net/publication/301355727_White_Grey_and_Black-Box_Modelling_in_Ship_Performance_Evaluation.

Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: Data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman*. Springer series in statistics. Springer, New York, 2nd ed. edition, 2009. ISBN 9780387848570. doi:10.1007/b94608.

U. Hollenbach. Estimating resistance and propulsion for single-screw and twin-screw ships in the preliminary design. In Chryssostomos Chryssostomidis and Kaj. Ed Johansson, editors, *10th international conference on computer applications in shipbuilding*, International conference on computer applications in shipbuilding, pages 237–250. 1999. ISBN 1561720240.

Leo H. Holthuijsen. *Waves in oceanic and coastal waters*. Cambridge University Press, Cambridge, 2007. ISBN 9780521860284.

J. Holtrop. A statistical re-analysis of resistance and propulsion data. *Published in International Shipbuilding Progress, ISP, Volume 31, Number 363*, 1984. URL https://repository.tudelft.nl/islandora/object/uuid%3Aca12a502-fc85-45e4-a078-db7284127e3c.

J. Holtrop and G.G.J. Mennen. A statistical power prediction method. *Netherlands Ship Model Basin, NSMB, Wageningen, Publication No. 603, Published in: International Shipbuilding Progress, ISP, Volume 25, Number 290, October 1978*, 1978. URL https://repository.tudelft.nl/islandora/object/uuid%3A62c40df8-18cc-4225-a65a-54ff5c1609fb.

J. Holtrop and G.G.J. Mennen. An approximate power prediction method. *Netherlands Ship Model Basin, NSMB, Wageningen, Publication No. 689, Published in: International Shipbuilding Progress, ISP, Volume 29, Nr 335, 1982*, 1982. URL https://repository.tudelft.nl/islandora/object/uuid%3Aee370fed-4b4f-4a70-af77-e14c3e692fd4.

IMO. Revised guidelines for the onboard operational use of shipborne Automatic Identification Systems (AIS), 2015. URL https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx.

IMO. Fourth IMO GHG study 2020. *International Maritime Organization London, UK*, 2020.

K. Torsethaugen, S. Haver, and S. Norway. Simplified double peak spectral model for ocean waves. 2004. URL https://www.semanticscholar.org/paper/Simplified-Double-Peak-Spectral-Model-For-Ocean-Torsethaugen-Haver/0f1b1509791d441861ff6c2940dd13b1f939f149.

Seong-Hoon Kim, Myung-Il Roh, Min-Jae Oh, Sung-Woo Park, and In-Il Kim. Estimation of ship operational efficiency from ais data using big data technology. *International Journal of Naval Architecture and Ocean Engineering*, 12:440–454, 2020. ISSN 2092-6782. doi:10.1016/j.ijnaoe.2020.03.007. URL https://www.sciencedirect.com/science/article/pii/S2092678220300091.

Hans Otto Kristensen and Marie Lützen. Prediction of resistance and propulsion power of ships. *Clean Shipping Currents*, 1(6):1–52, 2012. ISSN 2242-9794.

Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer, New York, 2013. ISBN 9781461468486.

Xiao Lang. *Development of Speed-power Performance Models for Ship Voyage Optimization*. PhD thesis, 2020. URL https://www.researchgate.net/publication/347977212_Development_of_Speed-power_Performance_Models_for_Ship_Voyage_Optimization.

Xiaohe Li, Yuquan Du, Yanyu Chen, Son Nguyen, Wei Zhang, Alessandro Schönborn, and Zhuo Sun. Data fusion and machine learning for ship fuel efficiency modeling: Part i – voyage report data and meteorological data. *Communications in Transportation Research*, 2:100074, 2022. ISSN 27724247. doi:10.1016/j.commtr.2022.100074.

Anthony F. Molland. *The maritime engineering reference book: A guide to ship design, construction and operation / edited by Anthony F. Molland*. Butterworth-Heinemann, 2011. ISBN 9780080560090. doi:10.1016/B978-0-7506-8987-8.X0001-7.

Ulrik D. Nielsen and Jesper Dietz. Ocean wave spectrum estimation using measured vessel motions from an in-service container ship. *Marine Structures*, 69:102682, 2020. ISSN 09518339. doi:10.1016/j.marstruc.2019.102682.

Joan P. Petersen, Ole Winther, and Daniel J. Jacobsen. A machine-learning approach to predict main energy consumption under realistic operational conditions. *Ship Technology Research*, 59(1):64–72, 2012a. ISSN 0937-7255. doi:10.1179/str.2012.59.1.007.

Jóan Petur Petersen. Mining of ship operation data for energy conservation. 2011. URL https://orbit.dtu.dk/en/publications/mining-of-ship-operation-data-for-energy-conservation.

Jóan Petur Petersen, Daniel J. Jacobsen, and Ole Winther. Statistical modelling for ship propulsion efficiency. *Journal of Marine Science and Technology*, 17(1): 30–39, 2012b. ISSN 0948-4280. doi:10.1007/s00773-011-0151-0.

Stian Glomvik Rakke. *Ship emissions calculation from AIS*. PhD thesis, NTNU, 2016. URL https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2410741.

D. Ronen. The effect of oil price on containership speed and fleet size. *Journal of the Operational Research Society*, 62(1):211–216, 2011. ISSN 0160-5682. doi:10.1057/jors.2009.169.

H. Schneekluth and Volker Bertram. *Ship design for efficiency and economy*. Butterworth-Heinemann, Oxford, 2nd ed. edition, 1998. ISBN 9780080517100. URL https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=313756.

T.W.P. Smith, J.P. Jalkanen, B.A. Anderson, J. J. Corbett, J. Faber, S. Hanayama, E. O'Keeffe, S. Parker, L. Johansson, L. Aldous, C. Raucci, M. Traut, S. Ettinger, D. Nelissen, D. S. Lee, S. Ng, A. Agrawal, J. J. Winebrake, M. Hoen, S. Chesworth, and A. Pandey. Third imo greenhouse gas study 2014. 2015. URL https://research.manchester.ac.uk/en/publications/third-imo-greenhouse-gas-study-2014.

Omer Soner, Emre Akyuz, and Metin Celik. Use of tree based methods in ship performance monitoring under operating conditions. *Ocean Engineering*, 166: 302–310, 2018. ISSN 00298018. doi:10.1016/j.oceaneng.2018.07.061. URL https://www.sciencedirect.com/science/article/pii/S0029801818314446.

Stopford. The organization of the shipping market. page 47, 2009.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pages 278–282 vol.1, 1995. doi:10.1109/ICDAR.1995.598994.

Shuaian Wang and Qiang Meng. Sailing speed optimization for container ships in a liner shipping network. *Transportation Research Part E: Logistics and Transportation Review*, 48(3):701–714, 2012. ISSN 13665545. doi:10.1016/j.tre.2011.12.003. URL https://www.sciencedirect.com/science/article/pii/S1366554511001554.

N. Wijnolst, Tor Wergeland, and Kai Levander. *Shipping Innovation*. IOS Press, 2009. ISBN 9781586039431.

Ran Yan, Shuaian Wang, and Yuquan Du. Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship. *Transportation Research Part E: Logistics and Transportation Review*, 138:101930, 2020. ISSN 13665545. doi:10.1016/j.tre.2020.101930.

Ran Yan, Shuaian Wang, and Harilaos N. Psaraftis. Data analytics for fuel consumption management in maritime transportation: Status and perspectives. *Transportation Research Part E: Logistics and Transportation Review*, 155:102489, 2021. ISSN 13665545. doi:10.1016/j.tre.2021.102489. URL https://www.sciencedirect.com/science/article/pii/S1366554521002519.

Dong Yang, Lingxiao Wu, Shuaian Wang, Haiying Jia, and Kevin X. Li. How big data enriches maritime research – a critical review of automatic identification system

(ais) data applications. *Transport Reviews*, 39(6):755–773, 2019. ISSN 0144-1647. doi:10.1080/01441647.2019.1649315. URL https://www.researchgate.net/publication/334738291_How_big_data_enriches_maritime_research_-_a_critical_review_of_Automatic_Identification_System_AIS_data_applications.

Liqian Yang, Gang Chen, Jinlou Zhao, and Niels Gorm Malý Rytter. Ship speed optimization considering ocean currents to enhance environmental sustainability in maritime shipping. *Sustainability*, 12(9):3649, 2020. doi:10.3390/su12093649.

Yang Zhou, Winnie Daamen, Tiedo Vellinga, and Serge P. Hoogendoorn. Impacts of wind and current on ship behavior in ports and waterways: A quantitative analysis based on ais data. *Ocean Engineering*, 213:107774, 2020. ISSN 00298018. doi:10.1016/j.oceaneng.2020.107774.

Declaration in lieu of oath

I hereby solemnly declare that I have independently completed this work or, in the case of group work, the part of the work that I have marked accordingly. I have not made use of the unauthorised assistance of third parties. Furthermore, I have used only the stated sources or aids and I have referenced all statements (particularly quotations) that I have adopted from the sources I have used verbatim or in essence.
I declare that the version of the work I have submitted in digital form is identical to the printed copies submitted.
I am aware that, in the case of an examination offence, the relevant assessment will be marked as 'insufficient' (5.0). In addition, an examination offence may be punishable as an administrative offence (Ordnungswidrigkeit) with a fine of up to €50,000. In cases of multiple or otherwise serious examination offences, I may also be removed from the register of students.
I am aware that the examiner and/or the Examination Board may use relevant software or other electronic aids in order to establish an examination offence has occurred
I solemnly declare that I have made the previous statements to the best of my knowledge and belief and that these statements are true and I have not concealed anything.
I am aware of the potential punishments for a false declaration in lieu of oath and in particular of the penalties set out in Sections 156 and 161 of the German Criminal Code (Strafgesetzbuch; StGB), which I have been specifically referred to.

Section 156 False declaration in lieu of an oath

Whoever falsely makes a declaration in lieu of an oath before an authority which is competent to administer such declarations or falsely testifies whilst referring to such a declaration incurs a penalty of imprisonment for a term not exceeding three years or a fine.

Section 161 Negligent false oath; negligent false declaration in lieu of oath

(1) Whoever commits one of the offences referred to in Sections 154 to 156 by negligence incurs a penalty of imprisonment for a term not exceeding one year or a fine. (2) No penalty is incurred if the offender corrects the false statement in time. The provisions of Section 158 (2) and (3) apply accordingly.

Place,date

Signature