

## 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:    | Prof. Dr. Ucker                    |
| Studiengang:     | ISE General Mechanical Engineering |
| Studiensemester: | Summer semester 2023               |
| Datum:           | 04.05.2023                         |

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                     | <b>5</b>  |
| <b>2</b> | <b>Theoretical Background</b>           | <b>6</b>  |
| 2.1      | Literature Review . . . . .             | 6         |
| 2.2      | Random Forest Regressor (RFR) . . . . . | 7         |
| 2.3      | Ship speed . . . . .                    | 7         |
| 2.4      | Modelling . . . . .                     | 7         |
| <b>3</b> | <b>Research Methodology</b>             | <b>8</b>  |
| 3.1      | Data Preprocessing . . . . .            | 8         |
| 3.2      | Data Analysis . . . . .                 | 8         |
| 3.3      | Modelling . . . . .                     | 11        |
| 3.4      | Predicting STW . . . . .                | 12        |
| <b>4</b> | <b>Result and Discussion</b>            | <b>14</b> |
| 4.1      | Model Evaluation . . . . .              | 14        |
| <b>5</b> | <b>Summary and Outlook</b>              | <b>16</b> |
|          | <b>References</b>                       | <b>17</b> |

List of Tables

|   |                             |    |
|---|-----------------------------|----|
| 1 | Model performance . . . . . | 14 |
| 2 | Model performance . . . . . | 14 |

List of Figures

|   |                                      |    |
|---|--------------------------------------|----|
| 1 | Example of partition space . . . . . | 7  |
| 2 | Example of partition tree . . . . .  | 7  |
| 3 | Histogram of the features . . . . .  | 9  |
| 4 | Correlation Heat Map . . . . .       | 11 |
| 5 | Correlation Heat Map . . . . .       | 15 |

## 1 Introduction

The research on a more efficient operation of vessel is a direction that is being actively pursued in the maritime world. Efficient ship operation translates to a noticeable increase in profitability and sustainability. The Fuel Oil Consumption (FOC) is a determining factor in the ship's operating cost, with fuel cost making up to 75% of the total ship operating cost [1, 2]. Different ship parameters affect the ship's fuel consumption and research shows that the reduction of ship speed is the most economical method to reduce fuel consumption. Precise forecasting and modelling of the ship speed will lead to a significant reduction in a ship's operating cost and could possibly extend the longevity of the ship.

In this data-driven thesis, powerful machine learning method will be utilized for modelling and forecasting of the ship's speed. To develop the model. The data used will be from the voyage of a Hammerhus RoPax ship between the port of Koege and Roenne. The modelling and forecasting using Random Forest Regressor will be emphasis of this thesis. The developed model of the ship's speed is then to be validated and where applicable, further optimized to increase its performance. This model is then to be used to predict the ship's fuel consumption.

## 2 Theoretical Background

This chapter deals with the past and present research in the relevant area which include literature review. This includes the significance of precise modelling of the ship's speed and its subsequent use in forecasting the ship's operation. The theoretical background of Random Forest Regression will be discussed in this chapter

### 2.1 Literature Review

The research to estimate ship fuel consumption is an active field of research. The summary of previous research compiled by Kim et al. [3] gives a good overview of the methodology used to estimate Fuel Oil Consumption (FOC). For example, Beşikçi et al. [4], Jeon et al. [5] and Kim et al. [3] utilised ANN to estimate FOC and reported good results by comparing ANN model's performance against Multiple Linear Regression model. It will be also sensible to compare the performance of ANN model against other machine learning models.

The research by Gkerekos et al. [2] showcased the performance of different machine learning models to predict ship's (FOC) using both noon data and automated data logging and monitoring (ADLM) system. This research concludes that Decision Tree Regressor (DTR) based model, namely Random Forest Regressor, (RFR), and Extra Tree Regressor, (ETR), displayed good prediction performance. Gkerekos et al. [2] also suggested that automatic sensor based data have the potential to increase the model accuracy score,  $R^2$ , by 5 – 7%.

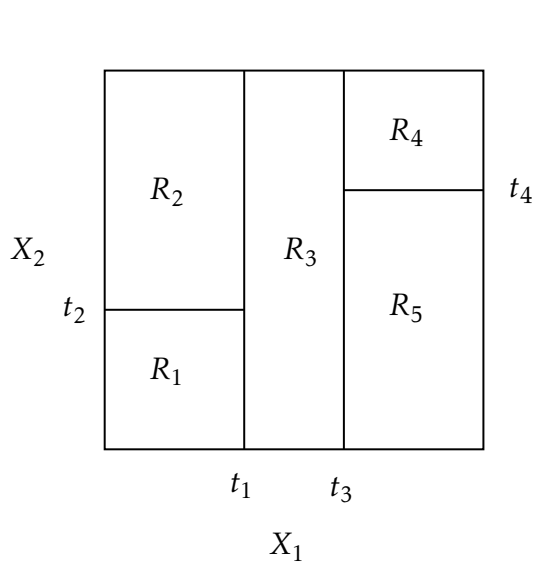
Li et al. [6] performed a more extensive research on the effects of data fusions between meteorological data, ship voyage data and AIS data on different machine learning models to predict the ship's FOC. This research highlighted the advantage of fusing meteorological data and ship voyage data. The evaluation on different model performance indicated that RFR are among preferable model candidate to be used in commercial scale due to its good prediction capability and robustness against different datasets.

The research presented by Gkerekos et al. [2] and Li et al. [6] both used machine learning method to estimate the ship's FOC. Abebe et al. [1] used different approach in their research by predicting the ship's Speed Over Ground (SOG) instead of FOC. Abebe et al. [1] fused AIS satellite data and noon-report weather data for the SOG prediction. The evaluation from this research shows that machine learning methods can also be applied to predict SOG. Findings from this research also show good prediction performance of RFR.

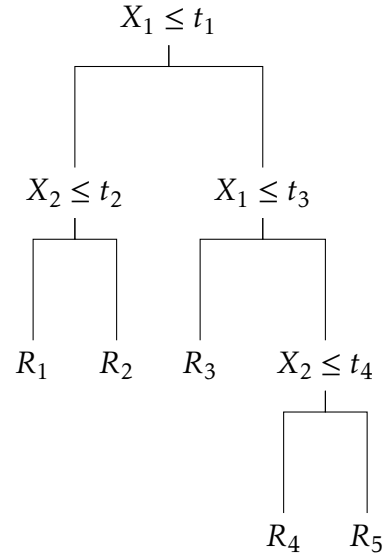
This literature review indicated the capability of Random Forest Regressor to predict SOG and FOC. However, the difference of the data source as well as the data type used for modelling will result different result in model performance. From the literature review, good prediction performance from RFR can be expected. With that,

this thesis will present the strength and limitations of RFR and provide general suggestions to improve the prediction performance of random forest model.

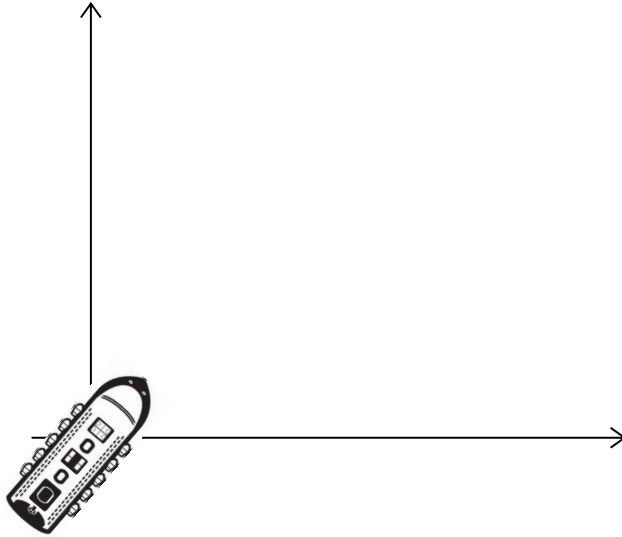
## 2.2 Random Forest Regressor (RFR)



**Figure 1:** Example of partition space



**Figure 2:** Example of partition tree



Suppose the following decision tree regressor

## 2.3 Ship speed

## 2.4 Modelling

## 3 Research Methodology

In this chapter the methodology used to develop the model will be discussed. The discussion on different parameters in the vessel's journey data will be discussed here. This includes the mining and merging of the features. The method used to develop the ship's speed model will be discussed in this chapter. This consists of the parameter used to develop the model. Ultimately, the model is then used to predict the ship's fuel consumption.

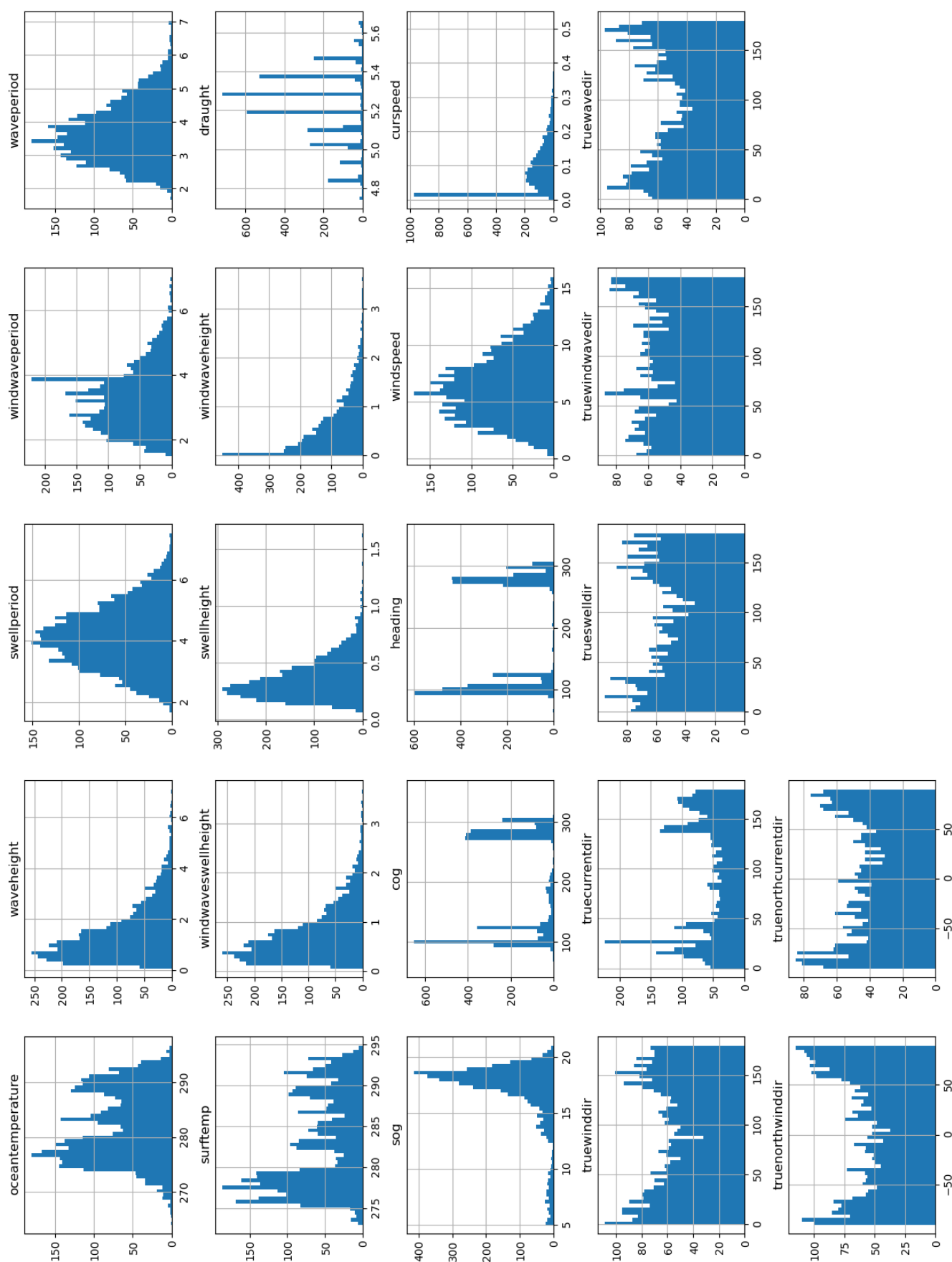
### 3.1 Data Preprocessing

- 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 [1,2,4,7]. This threshold is selected as 5 knots according to [1]
- The AIS data from June is filtered. This data will be used as validation data to check the model's performance.

### 3.2 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.





- 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, [1] 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.

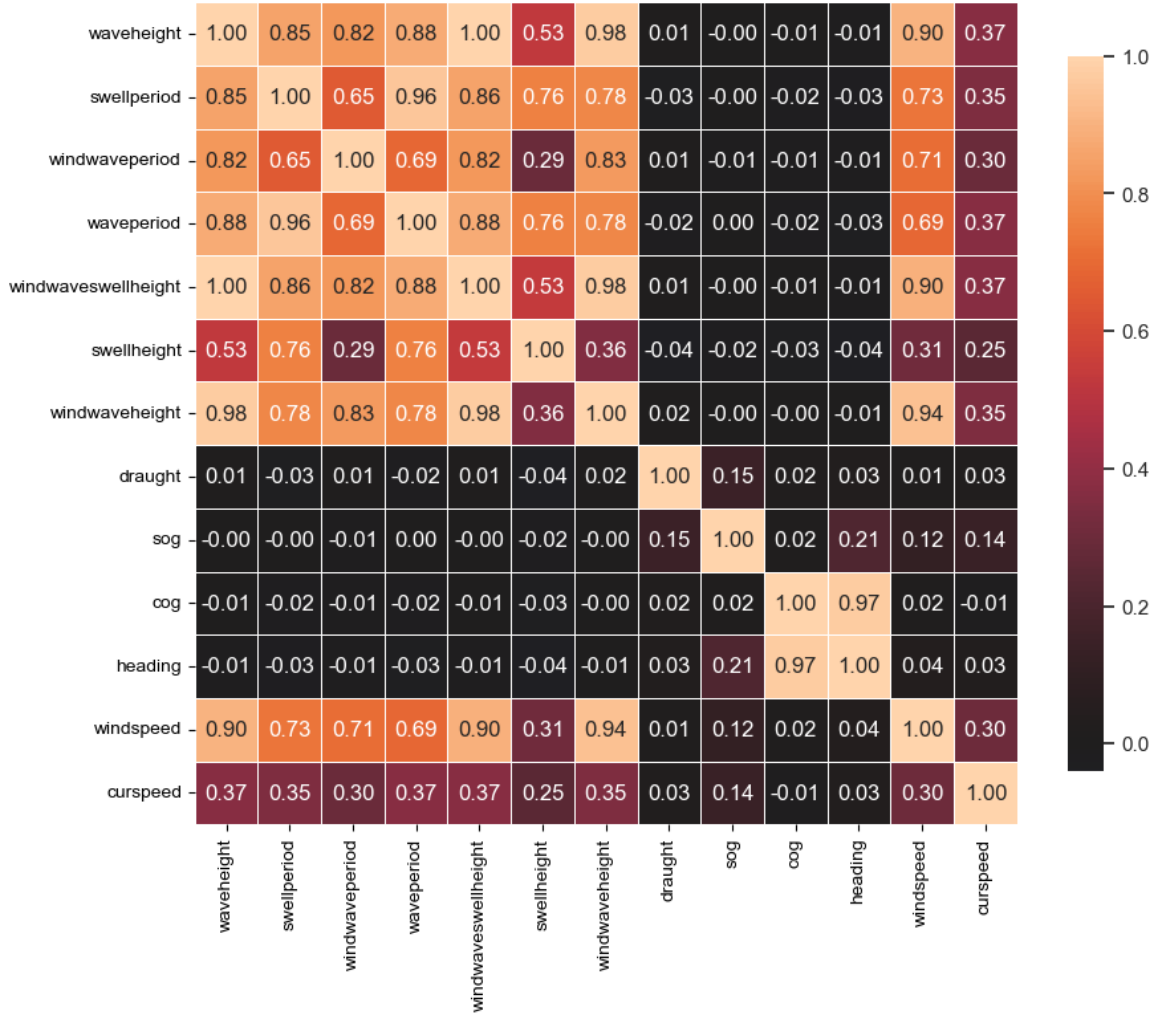


Figure 4: Correlation Heat Map

### 3.3 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

### 3.4 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. [7, 8]
- 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 [7]:

- 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)  $\beta$  with respect to True North:

$$V_g^x = V_g \sin(\beta) \quad (1)$$

$$V_g^y = V_g \cos(\beta) \quad (2)$$

- 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  $\gamma$  with respect to True North:

$$V_c^x = V_c \sin(\gamma) \quad (3)$$

$$V_c^y = V_c \cos(\gamma) \quad (4)$$

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

$$V_w^x = V_g^x - V_c^x \quad (5)$$

$$V_w^y = V_g^y - V_c^y \quad (6)$$

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

$$V_w = \sqrt{(V_w^x)^2 + (V_w^y)^2} \quad (7)$$

- This principle is applied to the following Python script. 3

```
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 reversion 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
```

## 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.

### 4.1 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 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 2:** Model performance

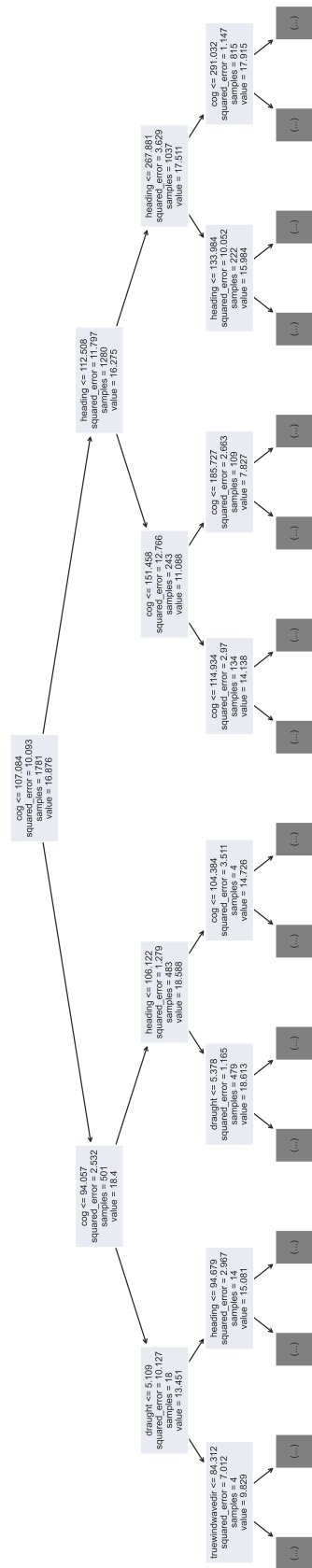


Figure 5: Correlation Heat Map

## 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.



## References

- [1] 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. 5, 6, 8, 10
- [2] 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. doi:10.1016/j.oceaneng.2019.106282. 5, 6, 8
- [3] Young-Rong Kim, Min Jung, and Jun-Bum Park. Development of a fuel consumption prediction model based on machine learning using ship in-service data. *Journal of Marine Science and Engineering*, 9(2):137, 2021. doi:10.3390/jmse9020137. 6
- [4] 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. doi:10.1016/j.cor.2015.04.004. 6, 8
- [5] Miyeon Jeon, Yoojeong Noh, Yongwoo Shin, O-Kaung Lim, Inwon Lee, and Daeseung Cho. Prediction of ship fuel consumption by using an artificial neural network. *Journal of Mechanical Science and Technology*, 32(12):5785–5796, 2018. URL: <https://link.springer.com/article/10.1007/s12206-018-1126-4>, doi:10.1007/s12206-018-1126-4. 6
- [6] Xiaohu 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. doi:10.1016/j.commtr.2022.100074. 6
- [7] 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. 8, 12
- [8] 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. doi:10.1016/j.oceaneng.2020.107774. 12

## 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