Journal of *Marine Science and Engineering*

MDPI

Article

Interpretable Machine Learning: A Case Study on Predicting Fuel Consumption in VLGC Ship Propulsion

Aleksandar Vorkapić [1,2,*], Sanda Martinčić-Ipšić [1,3,*] and Rok Piltaver [1]

[1] Faculty of Informatics and Digital Technologies, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia; rok.piltaver@inf.uniri.hr
[2] Faculty of Maritime Studies, University of Rijeka, Studentska 2, 51000 Rijeka, Croatia
[3] Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia
* Correspondence: a.vorkapic@icloud.com (A.V.); smarti@uniri.hr (S.M.-I.)

Abstract: The integration of machine learning (ML) in marine engineering has been increasingly subjected to stringent regulatory scrutiny. While environmental regulations aim to reduce harmful emissions and energy consumption, there is also a growing demand for the interpretability of ML models to ensure their reliability and adherence to safety standards. This research highlights the need to develop models that are both transparent and comprehensible to domain experts and regulatory bodies. This paper underscores the importance of transparency in machine learning through a use case involving a VLGC ship two-stroke propulsion engine. By adhering to the CRISP-DM standard, we fostered close collaboration between marine engineers and machine learning experts to circumvent the common pitfalls of automated ML. The methodology included comprehensive data exploration, cleaning, and verification, followed by feature selection and training of linear regression and decision tree models that are not only transparent but also highly interpretable. The linear model achieved an RMSE of 23.16 and an MRAE of 14.7%, while the accuracy of decision trees ranged between 96.4% and 97.69%. This study demonstrates that machine learning models for predicting propulsion engine fuel consumption can be interpretable, adhering to regulatory requirements, while still achieving adequate predictive performance.

Keywords: interpretability; machine learning; decision trees; linear regression; feature selection; two-stroke marine engines; fuel consumption

Citation: Vorkapić, A.; Martinčić-Ipšić, S.; Piltaver, R. Interpretable Machine Learning: A Case Study on Predicting Fuel Consumption in VLGC Ship Propulsion. *J. Mar. Sci. Eng.* **2024**, *12*, 1849. https://doi.org/10.3390/jmse12101849

Academic Editors: Lingxiao Wu and Shuaian Wang

Received: 12 September 2024
Revised: 13 October 2024
Accepted: 14 October 2024
Published: 16 October 2024

1. Introduction

Machine learning (ML) is impacting diverse maritime aspects by predicting operational parameters and addressing sustainability issues. By analyzing extensive data, including vessel performance metrics, weather, and environmental conditions, ML offers innovative solutions to enhance efficiency [1–4] and promote cleaner transport by reducing harmful emissions [5–7]. ML can also improve the resilience and robustness of onboard systems, contributing to the digital transformation of the maritime industry. This ongoing integration of ML allows for continuous adaptation to changing conditions, advancing maritime operations.

The development of ML models for predicting ship propulsion energy consumption represents an example of advancement in the marine industry [1,2,7]. Additional empirical and methodological research within the marine engineering sector includes performance optimization and knocking investigation into dual fuel two-stroke engines [1], regression models for predicting ship power [2], and monitoring operating behavior in propulsion diesel engines [7].

Recent advancements have extended the research and development of ML models, tasked with the enhancement of ship lifecycle management and operational efficiency. Nielsen et al. [3] enhance ship-maneuvering prediction by integrating a recurrent neural

network (RNN) with a first-principles model. Using data from full-scale ship recordings, the authors identify the limitations of conventional models in accurately predicting ship velocities, especially in confined waters and during maneuvers. They implement a hybrid approach where an RNN compensates for deviations between the measured velocities and the first-principles model's output. This methodology significantly improves prediction accuracy for surge, sway, and yaw velocities, providing a more reliable tool for applications like simulator training and propulsion performance monitoring. Coraddu et al. [8] use a data-driven model to estimate speed loss due to marine fouling, demonstrating superior accuracy over traditional methods and contributing to reduced fuel consumption and maintenance needs. These studies underscore the transformative impact of ML in marine engineering, enhancing operational efficiency across various applications.

This study builds upon previous research demonstrating the predictive capabilities of ML in maritime engineering, from foundational theoretical approaches to practical applications in diverse settings [5,6,9]. The deployment of ML systems in complex applications has heightened interest in optimizing not only for performance but also transparency, safety, nondiscrimination, and decision interpretability [10]. The absence of these auxiliary criteria is particularly concerning in high-stakes environments such as navigational support and autonomous ships, where transparency, accountability, reliability, and safety are essential [10].

Furthermore, a consensus on the definition of interpretability in machine learning and how it should be evaluated for benchmarking purposes is still missing. Researchers have provided different definitions for interpretability in ML models, ranging from deductive–nomological approaches [11] to sense-making through mechanisms [12,13]. Nevertheless, all the proposed approaches refer to the extent to which a user can comprehend and explain the ML model outcomes [10,14]. Interpretability assessments typically fall into two categories: one evaluates the usefulness of the system in practical applications or simplified versions thereof to determine interpretability [10]. The other approach uses quantifiable proxies to argue that certain model classes, such as linear models, rule lists, or decision trees [15,16], are inherently interpretable and therefore preferable for scenarios requiring transparent decision-making. Decision trees offer inherent interpretability as they provide explicit rules for decision-making based on the input features [14–16] and linear regression is known for being one of the most interpretable ML models [8,14,16,17]. Hence, in this study we opt to investigate the second category of models in the real-life use case of analyzing operational data from two-stroke marine engines. By closely examining decision tree structures and evaluating the relationships between operational parameters, we seek to understand how these parameters contribute to the ML model that predicts engine performance as measured by fuel consumption.

This study aims to address the technical complexities of ML models while emphasizing the need for transparency and interpretability, which is essential for model validation. The conducted ML experiments highlight the importance of collaboration between domain experts in marine engineering and machine learning to enhance model interpretability. Additionally, understandable models not only aid in meeting the regulatory requirements but also facilitate in explaining predictions to the stakeholders [18,19]. Understanding and interpreting ML model decisions are essential for maintaining these qualities across different domains.

Our study demonstrates that machine learning models for predicting propulsion engine fuel consumption can be designed to be interpretable, i.e., complying with regulatory standards, while effectively addressing challenges in marine engineering by achieving adequate predictive performance. The key contributions of the paper include advancing the model interpretability for predicting fuel consumption in very large gas carrier (VLGC) ship propulsion, deepening the understanding of system behavior, employing diverse feature selection methods to optimize variable sets, and highlighting the necessity for multidisciplinary collaboration between domain and ML experts.

The paper is structured as follows: Section 2 overviews the related work, Section 3 provides materials and methods with used dataset, tools, methods, and modeling process; Section 4 validates the ML methods and provides interpretability discussion, followed by Section 5, which concludes the paper.

2. Related Work

Machine learning, a branch of artificial intelligence, develops algorithms that learn from data to make predictions or decisions. These algorithms train models on specific datasets to automate and enhance decision-making processes. Data mining (DM) focuses on extracting useful information and patterns from datasets, employing ML algorithms to predict trends and inform decisions, using techniques like clustering and classification. Throughout the paper, both terms are used interchangeably: ML is referred to when discussing algorithms; and DM is used when performing data analyses [20–24]. Specifically, insights gained from DM are used to define the problem and prepare the data for ML algorithms.

Recently, interpretable machine learning (ML) models have gained significant attention in the ML literature [10,14,15,20,25–27]. In critical applications such as medicine [13], self-driving cars [10], and ships [10], understanding model reasoning (i.e., how the model makes decisions) is as crucial as performance [26]. The transparency of the model [20,26] is essential for comprehending the model structure (e.g., decision trees [14,15]), understanding individual components (e.g., parameters in regression [16,25]), and assessing the contribution of each input variable to the prediction outcome (e.g., feature importance [28–30]). The need for the interpretability of ML models has been emphasized by EU legislation. The General Data Protection Regulation (GDPR) has, since 2018, granted citizens the right to an explanation if they are affected by algorithmic decision-making [18]. Starting in 2024, the AI ACT, an EU regulation for artificial intelligence, establishes stringent requirements for the transparency of ML models used [19]. It has been shown that interpretability is inherent to so-called transparent ML models, while state-of-the-art solutions for opaque "black box" models are still a ML research challenge [10,20,26,27]. Several studies [1,2,10] on applied ML methods for marine engineering problems emphasize the drawbacks of utilizing opaque ML methods.

Kim et al. [2] employ a multilayer perceptron (MLP) to estimate ship power by modeling the performance characteristics of the hull form, focusing on resistance, propulsion, and propeller open water (POW) characteristics. Additionally, Convolutional Neural Networks (CNNs) are used to interpret the hull's geometry from images to forecast ship hydrodynamics. These prediction models are deemed suitable for the early design phase, particularly where a CNN model narrows down the selection of hull-form options. However, the authors caution that "data-driven prediction models, like the ones discussed, should be approached with caution for entirely new hull shapes not covered in the training data". This caution stems from the black-box nature of the utilized MLP and CNN, which do not provide explanations for their outputs or reasons for preferring one hull shape over another.

In the study by Jin et al. [1], a response surface model (RSM) is integrated with multi-objective particle swarm optimization (MOPSO) to enhance various parameters of the engine, with the goals of boosting overall performance and minimizing engine combustion. The authors note "The 1D simulation carried out provides good numerical simulation research on engine operating parameters, but it cannot clearly obtain the development of the combustion flame in the cylinder, at the same time, the parameters that can affect the knocking characteristics such as vortex and turbulence in the cylinder cannot be studied". This limitation also underscores the black-box characteristics inherent to the applied ML methods.

Campos et al., in [10], train a random forest (RF) classification model on merged forecast data to determine the wave height. The study examines the variable impact on the RF model accuracy, and authors retain the wave height, direction (swell), and period. Hence, this ablation study, inherent to RF classifier, explains the trained model. Still, the

authors report that the model achieves satisfactory performance for shorter periods but lags for longer ones, so extension with satellite images and training of additional ML methods is needed to improve the long-term prediction results.

Our study extends beyond the reported work in marine engineering ML by providing guidelines for greater transparency within the standard framework for creating ML models. We underscore not only the interdependencies of variables [5], but also the importance of feature selection, model complexity, performance, and interpretability in a marine engineering ML context. The conducted experiments aim to demonstrate that feature reduction can improve the robustness and interpretability of a model, while utilizing all available features may enhance predictive accuracy, potentially at the expense of increased model complexity. This approach provides a deeper insight into the relationships among variables, transcending traditional interdependence analysis found in previous works [4]. The results demonstrate that the choice between these approaches should be guided by the specific requirements of the application, balancing the demands for accuracy, complexity, and interpretability.

3. Materials and Methods

In this study, sensory data from a very large gas carrier (VLGC), similar to the recent series of ships built by a South Korean shipbuilder, were utilized. The vessel has a capacity of 54,340 DWT, a length of 225 m, and a width of 37 m. The main engine of the ship is a two-stroke marine diesel engine with one turbocharger unit, providing a maximum output power of 12,400 kW. This power calculation considers a 15% sea margin and a 10% engine margin for factors such as fouled ship hull and heavy weather, ensuring a guaranteed speed of 16.8 knots at the design draft. The propulsion system consists of a single four-blade fixed-pitch propeller with a diameter of 7400 mm, directly connected to the main engine via a shafting system.

3.1. Data Source

For the analysis, sensory data were collected from the main propulsion diesel engine (MAN B&W 6G60ME-C9.2) automation system capturing parameters such as revolutions per minute (rpm) and other engine-related variables. Additionally, data from Kongsberg's K-Chief 600 alarm, monitoring, and control system provide information on temperature and other system-related variables. The 7 Hz sampling frequency was chosen as the densest interval available for capturing performance data.

The cylinder pressure is measured by the Kistler's 6613EQ13-C online combustion control piezoelectric sensors, which were directly mounted at each cylinder indicator cock. These online sensors were calibrated according to the requirements stated in the IMO NOx regulations and the manufacturer's recommendation. Fuel oil mass flow (output variable) was measured using Endress + Hauser's Proline Promass 80 Coriolis Mass Flow Measuring System, which meets the ISO 11631:1998 standard [31] with a total error of 0.15%. Two flow meters of the same type were installed, one at the engine fuel inlet and the other at the fuel outlet line, and the difference between the readings represents the consumed fuel oil. The shaft power is measured by MetaPower's torque meter, while temperature sensors were integrated into the K-Chief 600 system. During data collection, readings were taken at different engine speeds: 89 min^{-1} (representing normal continuous rating, NCR), 85 min^{-1} (requested speed setting during sailing), and 75 min^{-1}. The loads varied between 5712 kW and 10,164 kW, as measured at the shaft. To ensure measurement repeatability and comparability, the engine outlet cooling water temperature was automatically controlled at 89 °C, while the engine outlet lubricating oil temperature was controlled between 45 °C and 47 °C using temperature controllers. The engine is supplied by fuel oil, compliant to ISO 8217 standards [32], with net specific energy 40.33 MJ/kg. In total, 1018 data sample instances were collected. The variables with the abbreviation tags and measurement units are listed in Table 1.

Table 1. The list of variables with tag abbreviation, description, and measurement unit (obtained from sensor data).

Tag	Description	Unit	Tag	Description	Unit
rpm	Shaft revolutions	min^{-1}	pi1	Indicated Mean Eff. Press. cyl. 1	bar
pwr	Shaft power	kW	pi2	Indicated Mean Eff. Press. cyl. 2	bar
pComp1	Compression Pressure, cyl. 1	bar	pi3	Indicated Mean Eff. Press. cyl. 3	bar
pComp2	Compression Pressure, cyl. 2	bar	pi4	Indicated Mean Eff. Press. cyl. 4	bar
pComp3	Compression Pressure, cyl. 3	bar	pi5	Indicated Mean Eff. Press. cyl. 5	bar
pComp4	Compression Pressure, cyl. 4	bar	pi6	Indicated Mean Eff. Press. cyl. 6	bar
pComp5	Compression Pressure, cyl. 5	bar	piAvg	Indicated Mean Eff. Press. mean	bar
pComp6	Compression Pressure, cyl. 6	bar	slip	Apparent slip ratio	%
pAvg	Compression Pressure mean	bar	temp	Ambient Air Temperature	°C
pMax1	Firing pressure, cyl. 1	bar	tExhGas1	ME Exhaust gas temperature, cyl. 1 (MA007)	°C
pMax2	Firing pressure, cyl. 2	bar	tExhGas2	ME Exhaust gas temperature, cyl. 2 (MA008)	°C
pMax3	Firing pressure, cyl. 3	bar	tExhGas3	ME Exhaust gas temperature, cyl. 3 (MA009)	°C
pMax4	Firing pressure, cyl. 4	bar	tExhGas4	ME Exhaust gas temperature, cyl. 4 (MA010)	°C
pMax5	Firing pressure, cyl. 5	bar	tExhGas5	ME Exhaust gas temperature, cyl. 5 (MA011)	°C
pMax6	Firing pressure, cyl. 6	bar	tExhGas6	ME Exhaust gas temperature, cyl. 6 (MA012)	°C
pMaxAvg	Firing pressure mean	bar	fuel	ME Fuel consumption	kg/h

Note: Firing pressure refers to the peak cylinder pressure during the combustion cycle, commonly known as maximal indicated pressure.

3.2. Data Preparation

Data are logged across 7 days, 4 of which have 160–400 data points while the other 3 days have limited number of measurements (4 or 10 data points each). The vessel was subjected to varying daily operational conditions, resulting in notable differences in observed engine parameters when data for different dates are compared. One of the dates has lower values for all engine parameters (12–34% relative difference depending on the parameter), while the values for other dates do not differ significantly (within $\pm$12%). These lower values are attributed to setting the engine revolution speed to an economic fuel consumption mode (75 rpm), which is empirically above the scavenge air pressure threshold for engaging auxiliary blowers. This operational mode was confirmed during the onboard measurement process. Additionally, the parameter values on this day remained quite consistent due to steady engine operation throughout the measurement period, resulting in narrow distributions for each parameter compared to the variations observed on other days (see Appendix C for details).

The parameters of compression pressure, peak cylinder pressure, indicated mean effective pressure, and exhaust gas temperature are represented with seven variables each—one for measurement at each of the six engine cylinders and the seventh as the average over all cylinders. The pressure correlations for each pair of cylinders on a given day are higher than 0.97 in all cases and in most cases even higher than 0.99. Values at some cylinders are systematically higher or lower than on the other cylinders; however, the order of average values per cylinder changes with the dates. The observed cylinder pressure variations remain within the engine manufacturer's acceptable margins, reflecting the engine control system auto-balance and maintaining the set engine speed and performance parameters across a range of operational conditions. The exhaust temperatures at different cylinders are highly correlated (0.73–0.99) as well but not as much as the pressures. The ambient air temperature remains constant during each measurement interval because these periods are short, typically less than an hour. Like other parameters, fuel consumption is also constant on each day but differs between days, e.g., the day with the lowest average consumption has 43% lower value than the day with the highest consumption. The discretization of fuel consumption into nine buckets used as the class is described in Section 3.4. Discretized fuel consumption falls into a single or two adjacent bucket-classes for all days, but one that has fuel consumption spread over the top four narrow buckets.

The dataset, which comprises operational data from two-stroke marine engines, has been prepared for analysis. The measurement data from various sources were time-synchronized to ensure accurate alignment. These steps ensure that the data are appropriately processed and ready for further investigation.

3.3. Methods

The first ML method used to model the relationship between fuel consumption and the engine parameters is linear regression, which is a fundamental statistical method often used in ML to explain or predict a continuous value based on values of other parameters [16,25]. The concept behind linear regression is relatively straightforward: it assumes a linear relationship between the independent variables (engine parameters) and the dependent variable (fuel consumption) and finds the linear function that is the best possible fit, i.e., the straight line that minimizes the differences between the observed values and the values predicted by the model [21,22]. This line then serves as a predictive tool, allowing us to estimate fuel consumption based on the other engine parameters.

Multiple linear regression (MLR) extends the concept of simple linear regression to incorporate multiple independent variables (features) in predicting a dependent variable, providing a more comprehensive understanding of the relationship between the variables involved. Linear regression and evaluation metrics R-squared (R2), root mean squared error (RMSE), mean relative absolute error (MRAE), and mean absolute error (MAE) for regression models are elaborated in Appendix A.1.

The second utilized ML method is the decision tree method [1,15,24]. Decision tree is a non-parametric classification method widely applicable in various problem domains, including predictive supervised learning. One of the main advantages of this method is its interpretability, simplicity, and fast learning process. The algorithm constructs a tree structure from input variables, enabling straightforward analysis and interpretation of predictions [33,34]. Each node in the tree divides the input variable into child nodes for each variable value of the input variable in the parent node. For continuous variables, divisions are based on comparisons of values within specific intervals, while for discrete variables, divisions are based on combinations of all possible non-repeating values. Each leaf node in the tree represents the value of the target variable given the input variable values represented by the path from the root to the leaf. The tree grows by recursively partitioning the original dataset into subsets based on testing the variable values. The partitioning process continues until all data instances in a specific node have the same value as the target variable or when further branching does not contribute to the accuracy of predictions. It is possible that, with the available input variables, complete purity is not achieved in a leaf node. The decision tree algorithm is presented in Appendix A.2.

Overall, decision trees provide a comprehensible and interpretable approach for predicting fuel consumption in marine two-stroke diesel engines. The J48 algorithm [33,34], specifically, offers interpretability, simplicity, and fast learning capabilities, making it a suitable choice for this study. Therefore, we decided to supplement the regression model with a classification approach (although the underlying problem is regression-based) to check whether sacrificing some accuracy could yield better comprehensibility and interpretability. This decision is aligned with the goals of our study, as our focus lies on the interpretability of ML models, particularly decision trees. The use of J48 in the analysis seeks to uncover the relationships and decision rules within the model that contribute to predicting fuel consumption in maritime two-stroke engines. This choice also facilitates the interpretation of results; the classes a: $\leq$1042 kg/h, b: >1042 and $\leq$1130 kg/h, c: >1130 and $\leq$1415 kg/h, d: >1415 and $\leq$1620 kg/h, e: >1620 and $\leq$1650 kg/h, f: >1650 and $\leq$1790 kg/h, g: >1790 and $\leq$1815 kg/h, h: >1815 and $\leq$1828 kg/h, i: >1828 kg/h are easily understood, as they are derived by converting expected operational daily values into hourly values. This approach aids in optimizing maritime operations by providing valuable insights into the factors influencing engine performance and fuel efficiency.

3.4. Experimental Design

Data mining experimental design follows the CRISP-DM (Cross Industry Standard Process for Data Mining) [23]. Our data mining objective is to predict fuel consumption using comprehensible machine learning models that treat sensor data as either numerical or categorical variables. Secondary objectives include the general analysis of sensor data, detection of engine performance anomalies, and careful examination of the interplay between variables, the complexity of ML models, their performance, and interpretability. DM activities are conducted following the CRISP-DM standard with an initial set of tools and DM techniques proposed: multiple linear regression and classification trees are trained on a potentially reduced set of variables according to the results of the feature selection process.

The second CRISP-DM phase was dedicated to data understanding including extensive data visualization (time-series, distribution, and summary metrics for each variable/sensor and day, scatterplots, correlations, and attribute clustering), verifying data quality, and discussions about the observed patterns and anomalies with the marine engineer. This phase is elaborated in Sections 3.1 and 3.2. It was noticed that some of the parameters are highly correlated to fuel consumption, as expected. The highest correlation with average daily fuel consumption was noticed with the average daily shaft power (0.951) and the ship's apparent slip ratio (0.925). Correlations with the pressures and exhaust temperature are also high (0.86–0.91). In Figure 1, the Spearman correlation between the reduced set of variables is reported, while all correlations are detailed in Appendix C Figure A1.

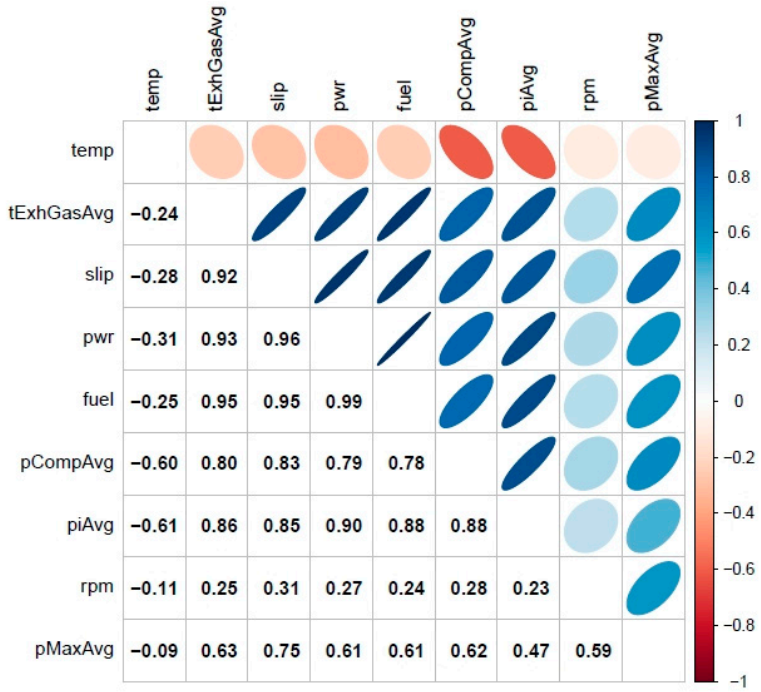

Figure 1. Spearman correlations between reduced set of variables.

The third CRISP-DM phase includes data preparation steps. Data from multiple sources were consolidated into a single dataset, duplicated data were removed, timestamps were interpolated based on sampling frequency (original data had 1 s accuracy, while sampling frequency was 420 min^{-1}), discretized fuel consumption was added as the class, and additional variables such as the date and the number of available data samples for the day were included to facilitate easier visualizations and validation. Variables were removed based on visualizations, the computed correlations between them, and discussion with the marine engineering expert. Our study prioritizes feature selection, model complexity, performance, and interpretability in the ML domain over the detection of operational anomalies or faults. Therefore, the dataset has been tailored to include only the aggregated

average values across all six cylinders, omitting the individual cylinder measurements such as compressor, firing, and indicated mean eff. pressures, as well as exhaust temperature. These detailed measurements are essential for engine health assessment but are not relevant to the aims of this study, which seeks to advance the strategic understanding of ML applications in marine engine performance. Date and time were also removed because all parameters including fuel consumption are very static; therefore, models tend to overfit by using the date to predict the fuel consumption. Ambient air temperature was removed due to the same reason—this decision was not initially obvious, but rather discovered during the data understanding phase.

Next, feature selection was performed. Three feature selection methods were applied to the complete and the reduced set of features were compared to demonstrate the importance of data understanding. It can only be achieved through careful data analysis and tight collaboration between domain and ML expert. However, once data are well understood, it becomes much easier to perform manual feature selection because it is well understood which features should be removed and why. Furthermore, it enables better feature engineering (adding new features), which can replace redundant features or transform existing features into features that help specific ML algorithms train better models.

The original plan was to use the common method based on correlation with the predicted variable. However, this method turned out to be unacceptable for the task due to high correlations between variable pairs (see Appendix C) and because this method cannot properly detect interactions between individual variables and remove the redundant ones. Therefore, alternative methods are used.

The first method is based on a genetic algorithm with a linear regression fitness evaluator [28] from the tidyfit R package that uses stochastic optimization to optimize the set of variables. This method produces different results compared to the correlation-based feature selection as is evident from Figure 2. It properly detected some important variables and marked the others as irrelevant (in combination with the top variables). However, it did not include shaft power as one of the top variables (although it has the high correlation with fuel consumption and is obviously the most important based on physics) and gave conflicting results especially when used on all features (for example giving positive weight to pressure on one cylinder but negative to the other cylinder). It was suspected that genetic algorithm-based feature selection suffered from overfitting due to the limited dataset and its optimization power.

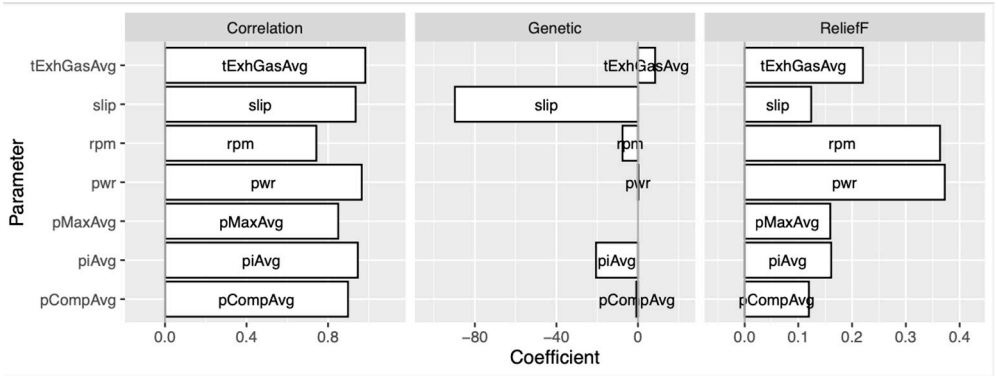

Figure 2. Feature importance for reduced set of variables according to correlation, genetic, and ReliefF methods.

Therefore, it was decided to use feature selection based on the third method ReliefF [28], which takes a filter–method approach and is notably sensitive to feature interactions. Again, we used the tidyfit R package that uses ReliefF implementation by [29]; that rated rpm and shaft power as the most important variables.

Note that manually removing redundant measurements at each individual cylinder helped all three feature selection algorithms obtain better results (Figure A3 in Appendix C).

Comparing the almost contradictory results of genetic and ReliefF-based algorithms shows that there are multiple small subsets of variables that can be used to train an accurate predictor, because all variables are highly correlated to fuel consumption due to their physical relation to fuel consumption. Finally, it was decided to proceed with feature selection based on both methods and rejecting feature selection based on correlation.

In the fourth CRISP-DM phase, regression models were trained using multiple linear regression (MLR) and classifiers were based on decision trees. Both methods were selected because they are renowned for their interpretability [14,15,26]. Linear regression models were trained using the lm function from the stats package that is based on [35] and decision trees using rpart function from rpart package which is based on [36].

Data instances were weighted so that instances for each of the 7 days contributed one-seventh of the total weight and all instances for the same date had the same weights. This prevents overfitting to days with many instances and ignoring days with few instances, which is important for our skewed dataset (see Section 3.1). Data were randomly split into training and testing sets with 80% of samples being used for training and 20% for testing. The split into training and testing set was stratified by date to ensure data for each date were present in both sets.

4. Results and Discussion

Next, the results obtained by linear regression and classification tree methods are reported and discussed, respectively.

Using all features (after manual feature selection) in a multiple linear regression model results in almost perfect model with R2 0.995, RMSE 19.79, MRAE 13.0%, and MAE 15.3. This is clearly an easy regression problem, as the variables are well correlated with the predicted variable and there are sufficient data. The *p*-values for all coefficients included in the regression model are almost zero except for shaft revolutions per minute and mean firing pressure, which are not significant (indicating that these parameters can be excluded from the model because they do not influence fuel consumption or because they are highly correlated with another parameter included in the model), and slip, which has a *p*-value of 0.012 (indicating that it is probably in a linear relationship with fuel consumption but not as strong as the other features).

Based on collaboration between the domain experts, the following model is obtained that is as accurate but simplified, making it easier to understand, validate, and apply:

$$\text{Fuel consumption} = -1756 + \text{avg. exhaust gas temperature} \times 8.639 + \text{shaft power} \times 0.2379 - \text{slip} \times 461.3 \quad (1)$$

This model achieved an RMSE of 23.16. Its average absolute error, 18.2 kg/h, means that the predictions are within 1–2% of the actual fuel consumption, which is between 1000 and 1700 kg/h. An MRAE of 14.7% means that the model is much more accurate than a simple model that would always predict an average fuel consumption. This is a satisfactory result, considering the limited dataset and the model's simplicity in terms of understanding and application.

The constant term (−1756) acts as a calibration offset, adjusting the baseline of the fuel consumption calculated from other parameters. Its negative value helps align the model outputs with observed data, where base fuel consumption levels are adjusted downwards. The positive coefficient associated with average exhaust temperature implies that an increase in exhaust temperature, indicative of higher power and potential heat losses, leads to increased fuel consumption. Shaft power relates to the engine's output or work performed. Its positive coefficient (0.2379) suggests that a higher power output requires more fuel. The negative coefficient for slip (−461.3) may seem counterintuitive because higher slip typically indicates less efficient propulsion. This anomaly could be due to insufficient data variability, confounding variables, or errors in model specification. The limited range of measured slip values (min. 2.86%, max. 4.18%, avg. 3.71%) restricts the ability to draw definitive conclusions about its effect on fuel consumption. With the presented simple multiple linear regression model, the prediction of fuel consumption is limited and may not

accurately reflect the true relationship between slip and fuel consumption. This limitation underscores the need for a broader dataset and a more sophisticated modeling approach to better understand and predict the effects of slip on fuel efficiency. The *p*-values for all coefficients included in this simplified regression model are almost zero, which means that according to this model all three parameters influence fuel consumption.

In addition to the empirical basis provided for the coefficients in our regression models, it is also possible to contextualize these coefficients within a theoretical framework based on the principles of thermodynamics, particularly relating to the efficiency and operation of marine engines. The basic theoretical model for engine performance can be derived from the first law of thermodynamics, which, in the context of an engine, is about the conversion of fuel energy into mechanical work and heat loss. The formula for the fuel consumption of an engine, which influences thermal efficiency, can be expressed as follows:

$$\mathbf{m_f} = \frac{\mathbf{P}}{\boldsymbol{\eta} * \mathbf{Q}} \tag{2}$$

Here, P represents the power output (kW), mf is the mass flow rate of fuel (kg/h), and Q is the heating value of the fuel (kJ/kg).

Thermal efficiency (η) of an engine can be further expressed as follows:

$$\eta = \eta_0 - C_1 \times \text{avg.exhaust gas temperature} - C_2 \times \text{slip}, \tag{3}$$

where η_0 represents baseline efficiency under standard operating conditions and C_1 and C_2 should be empirically dimensioned based on experimental data or detailed engine performance analyses.

The coefficients in our regression model, such as 8.639 and −461.3, can be linked to parameters in the thermodynamic efficiency equation. For instance, a coefficient related to power (P) in the model could be understood as reflecting changes in engine efficiency as power output varies. Similarly, negative coefficients might be associated with inefficiencies or increased fuel consumption due to factors like increased friction or thermal losses, as indicated by higher operational loads.

Bootstrapping was used to check the stability of linear regression model and calculate confidence intervals of predicted fuel consumption. Each linear regression model was fitted to 500 random samples from the training set and the process was repeated 1000 times. The models were then applied to the samples from the test set. The mean prediction value and the range that contains the predicted value for 95% of the 1000 models was computed for each sample in the test set. The results are shown in Figure 3, where the error bars can be interpreted as confidence in prediction or model stability depending on the subset of data used for training the model. The average width of the 95% confidence interval is 16.9, which amounts to less than 2% of the actual fuel consumption. Confidence intervals are narrower for test samples that have very low or very high actual fuel consumption, which means that the set of training data has lower impact on model predictions for such cases. Narrow confidence intervals confirm that predictions do not change much if a different subset of training examples is used to fit the model. The average RMSE over the 1000 models is 18.5 and the average MAE is 17.9, which is in line with errors computed for the single model that was trained on the entire training set (see Equation (1)).

The results of prediction models based on feature selection are listed in Table 2. All models (even with more than three variables) perform worse than the model based on the manual selection. When comparing models based on the feature selection, the differences when reducing from five to three variables are minor. However, models based on features selected with genetic algorithm highly outperform feature selection based on ReliefF. Models based on the genetic feature selection may be slightly overfitted (RMSE on test data is higher than on training data), while this is not the case for the models based on ReliefF feature selection. Reducing the number of features selected by the genetic algorithm to two

decreases R2 to 0.856, which is lower than in the case of the top two features selected using ReliefF (0.944).

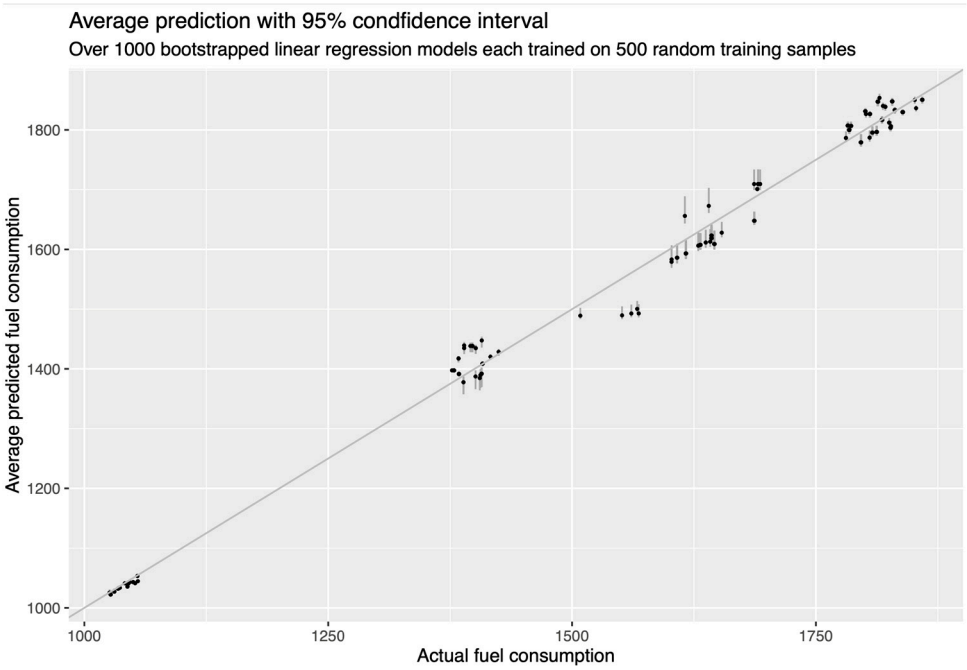

Figure 3. Dots represent the average predicted fuel consumption over the 1000 bootstrapped models for each sample in the test set. Error bars represent the range that contains 95% predictions. Distance from the diagonal line indicates the prediction error.

Table 2. Evaluation of regression models based on feature selection methods and number of variables used. Values in brackets are measured on training data.

Evaluation Metric	Genetic, Top 5 Attrib.	Genetic, Top 3 Attrib.	ReliefF, Top 5 Attrib.	ReliefF, Top 3 Attrib.	Manual, 3 Attrib.
R2	(0.987)	(0.967)	(0.955)	(0.954)	(0.992)
RMSE	29.04 (28.95)	27.42 (26.35)	57.04 (60.02)	59.52 (61.71)	23.16 (23.43)
MRAE [%]	19.2 (20.7)	13.8 (14.1%)	35.8 (37.3)	38.1 (39.9%)	14.7 (14.5%)
MAE [kg/h]	23.33 (24.98)	20.54 (21.81)	47.55 (48.74)	48.32 (49.83)	18.2 (18.45)

The best model based on feature selection uses three variables selected with the genetic algorithm:

$$\begin{aligned}\text{Fuel consumption} = -3005 + \text{slip} \times 49.87 + \text{avg. indicated mean eff. pressure} \\ \times\ 65.05 + \\ \text{avg. exhaust gas temperature} \times 10.02\end{aligned} \quad (4)$$

As discussed above, the constant term (-3005) is a calibration offset that adjusts the baseline of fuel consumption calculations based on the behavior observed in processed data. The positive coefficient for slip (49.87) is consistent with the physical principles and has a p-value of 0.0015. The coefficient for average indicated mean eff. pressure (65.05) suggests that higher pressures within the cylinders, indicative of more intense combustion processes, are associated with higher fuel consumption. Its p-value is almost 0 ($<2^{-16}$).

Next, a regression tree model is trained on dataset with all features after manual feature selection. This resulted in a binary tree with nine leaves (as many as classes) that uses shaft power as a splitting variable in most of the nodes and rpm and average pressure in only one node each. The tree is shown in Figure 4. The classification accuracy of the tree for 80:20 split training and testing sets is 96.4% (95.76% learned and tested on the complete

dataset) and the RMSE is 37.22. The RMSE of discretized fuel consumption (i.e., the actual values, not predictions) is 30.41, which means that most of the classification tree prediction error is due to the discretization of fuel consumption and not due to the prediction error. RMSE was computed by mapping each class's label to the mean fuel consumption over the examples belonging to the class.

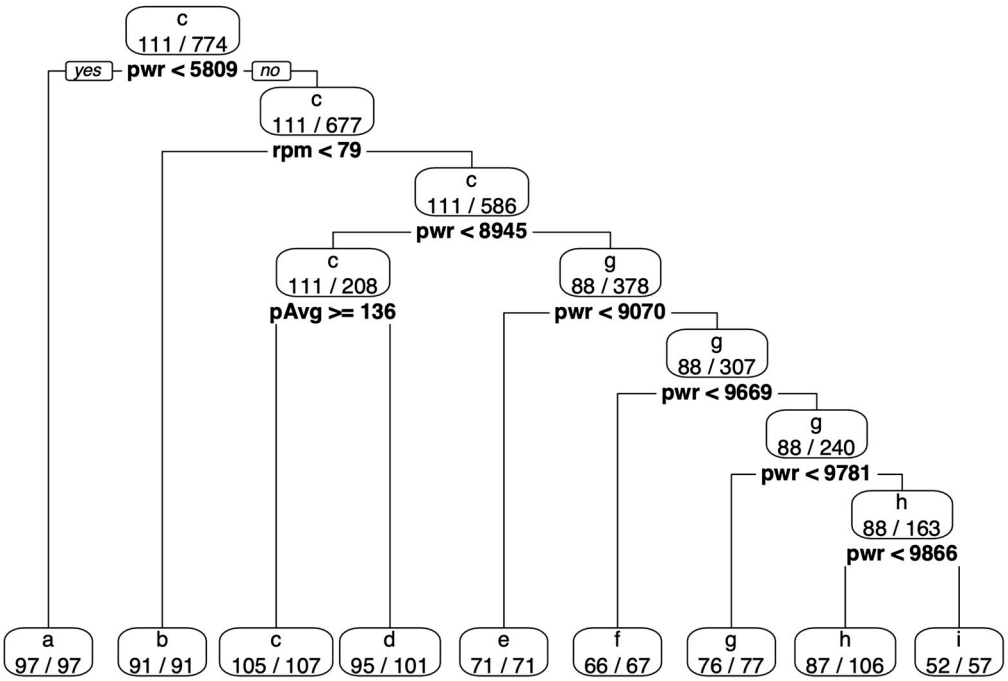

Figure 4. Classification tree that predicts fuel consumption classes a to i. Bold text represents the splitting criterion, letter represents the predicted class in each node and the numbers represent the number of training data samples belonging to the predicted class vs. all samples belonging to the node.

Most leaves are nearly pure, meaning they contain very few training examples from classes other than the predicted class. The purity of the leaves can be interpreted as an indication of confidence in the predictions—specifically, how confident the model is that the predicted class is correct. Class h has the lowest purity at 82.1%, while all other leaves exhibit over 90% purity; notably, the top six out of nine leaves have a purity exceeding 98%.

The tree can be simplified as follows: "the higher the shaft power the higher the fuel consumption, increasing rpm also increases consumption, at least when power is low". This result is in line with the obtained regression models and the theoretical model represented by Equation (2).

Detailed analysis of the decision tree's rules and node splitting reveals that shaft power is the primary variable for decision-making, showing a direct correlation between increased power output and higher fuel consumption. This relationship underscores a basic principle where greater energy output requires more fuel. Shaft revolutions per minute (rpm), although less frequently a criterion for splitting, significantly influences fuel consumption under specific conditions, particularly at lower power outputs where higher rpm leads to increased fuel usage. The indicated mean effective pressure (pAvg), while not a common splitting criterion, is critical under varying conditions and provides insights into the engine's combustion efficiency. The decision tree effectively uses pAvg to identify subtle variations in fuel consumption that are not readily evident through shaft power and rpm alone.

Classification trees were also trained on a dataset with feature selection applied using genetic and ReliefF methods described above. The trees based on the top five features according to ReliefF feature selection are exactly the same as the trees in Figure 4 (trained

using all features). The tree using only the top three features according to ReliefF feature selection is very similar, with the single difference of replacing the node split by pAvg with a subtree that substitutes a simple split by pAvg with two splits by pMax and two by pwr. Nevertheless, its performance is almost the same. On the other hand, trees based on top features according to the genetic feature selection method have more nodes (14 internal plus 15 leaves), which makes them more complex but not significantly more accurate.

Finally, simplified classification trees were trained using Weka software toolkit (ver. 3.8.6) to predict fuel consumption based on expert-based feature selection aimed for model simplicity. Specifically, rpm and shaft power were used, as suggested by the marine engineering expert. The decision trees were constructed with a standard implementation of the algorithms [22]. The J48 algorithm, which is Weka's implementation of C4.5 algorithm [33], was used. The goal of training simple classification trees is to validate the performance of these simple yet highly interpretable models.

The feature selection is based on the following reasoning and data interpretation. Firing pressure (pMax), compression pressure (pComp) and indicated mean eff. pressure (pi) from various cylinders are highly correlated with fuel consumption: the respective correlations for each group of parameters are ~0.85, ~0.9, and ~0.95 (see Appendix C). These parameters significantly impact the combustion process and thereby influence fuel consumption. Their relationship with exhaust gas temperature underscores the complex interaction between various engine parameters, dictated by thermodynamic principles. Moreover, slip also has a high correlation with fuel consumption at coefficients of 0.935. Slip can provide valuable insights into the performance of the hull and propeller, and the influence of environmental conditions such as waves, wind, and currents. These factors play a significant role in fuel efficiency.

Shaft revolutions (rpm), representing the number of engine cycles per minute, offer another essential piece of the puzzle. While a higher engine speed can suggest more fuel consumption due to the increased number of strokes, the relationship is not so direct. The actual load on the engine can vary under different operating and environmental conditions, potentially leading to higher fuel consumption at lower revolution speed under challenging conditions, and vice versa. This is reflected in the lower correlation with fuel consumption at coefficient of 0.743.

These findings highlight the interconnected nature of various operational parameters influencing fuel consumption. The observed correlations suggest that a more accurate prediction of fuel consumption may rely on a combination of these parameters rather than a single isolated factor. The analysis also points towards the need for a detailed investigation into parameters with lower correlation values to uncover potential performance issues and identify opportunities for optimization.

Hence, several classification trees are subsequently trained using J48/C4.5 decision tree algorithm, based on the expert insights. The first J48/C4.5 tree is trained on a single variable selected by the domain expert—the engine speed (rpm). Fuel consumption in marine engines is intricately linked to rpm, given its direct influence on power output. Therefore, the decision to focus on rpm in this first tree is aligned with the established understanding of engine operation. The resulting tree, as visualized in Figure 5, uses the rpm variable to split the data and predict the fuel consumption. Threshold values of rpm are used to create branches and leaves. The depth and complexity of the tree captures the nonlinear relationship between engine speed and fuel consumption.

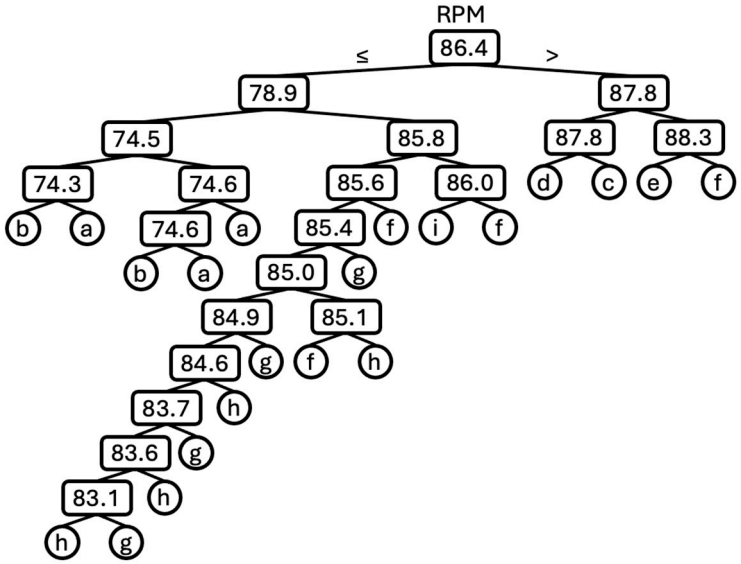

Figure 5. Pruned J48/C4.5 tree for shaft revolutions and fuel consumption variables and classes a to i.

It is well understood that an increase in rpm does not correspond to a linear increase in fuel consumption. Rather, the relationship may be exponential, with small increases in rpm leading to larger increases in fuel consumption, especially at higher rpm levels. This first model captures this intricate relationship with substantial fidelity, as evidenced by a high Kappa statistic of 0.6439 (Kappa statistics are defined in Appendix B). However, while the rpm is a major factor influencing fuel consumption, it is not the only one. The subsequent decision tree models incorporate expert-selected variables, such as shaft power, along with a comprehensive model that includes all available variables. The objective is to capture the most accurate and comprehensive understanding of the factors affecting fuel consumption in marine engines. While the above tree-based model provides valuable insights into the primary role of rpm in fuel consumption, the consequent models give a more holistic view of the system under study.

The second J48/C4.5 decision tree model shown in Figure 6, was constructed using only one parameter: shaft power (pwr). Just like the first model with rpm, this model illustrates that fuel consumption is influenced by the power output from the shaft. However, focusing more on the mechanical power fuel offers a different operational perspective. The model shows a high degree of accuracy, with a correct classification rate of 96.86%. The Kappa statistic of 0.9646 suggests good agreement between predictions and actual classes.

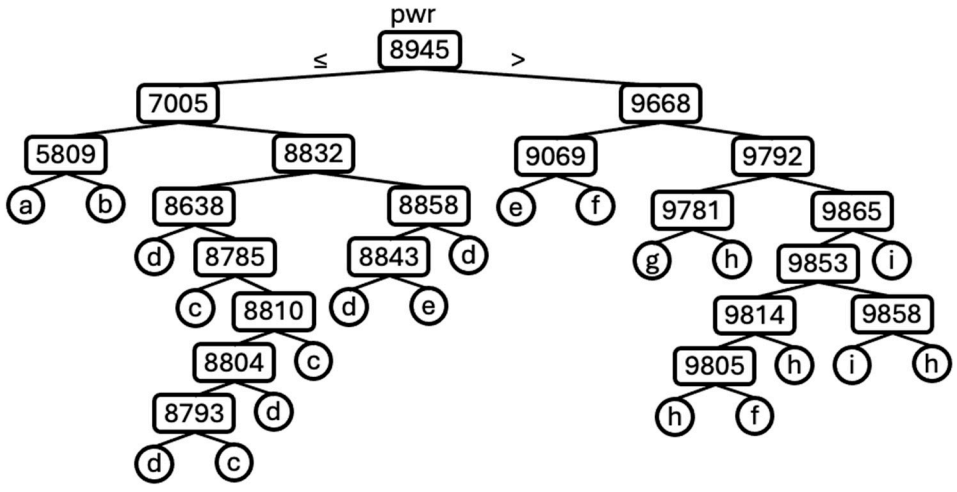

Figure 6. Pruned J48/C4.5 tree for shaft power and fuel consumption variables (classes a to i).

The detailed accuracy by class (reported in Table A4 in Appendix B) provides more insights into the model's performance. Each class, ranging from 'a' to 'i', corresponds to different intervals of shaft power. The model shows high true positive rates for all classes, signifying that most instances are classified correctly. The confusion matrix (see Table A5 in Appendix B) further confirms the model's superior performance. Most predictions align diagonally, which is the desired pattern, meaning that most instances were correctly classified.

The presented decision tree (Figure 6) explains how different shaft power ranges correspond to distinct fuel consumption classes. This model is slightly more complex, illustrating the nuanced relationship between shaft power and fuel consumption. For instance, the model suggests different fuel consumption classes within a narrow range of shaft power (e.g., between 8804.0 and 8810.0), indicating that changes in shaft power are not linearly following the fuel consumption.

The decision rules derived from this tree are straightforward and easy to interpret. For instance, if the shaft power is less than or equal to 8945 and further less than or equal to 7005, and within this subset it is less than or equal to 5809, the model predicts the fuel consumption class 'a'. The other rules can be interpreted similarly.

In conclusion, the second decision tree model using shaft power as the main predictor demonstrates high accuracy and provides a rich set of rules for predicting fuel consumption. However, like the first model, it is still a simplification. A more comprehensive model using all the available variables was constructed next.

The third J48/C4.5 decision tree model, when evaluated using all the variables of a dataset divided into 66% for training and 34% for testing, displayed particularly good performance with a high level of accuracy. The model correctly classified 97.69% of the instances. The high level of accuracy yielded a Kappa statistic of 0.9739, signifying that there is excellent agreement between the actual and predicted classes.

The detailed accuracy for each class was assessed and is reported in Table A6. This revealed that the model performed robustly across all classes. The true positive (TP) rates range from 0.897 for class 'g' to a perfect score of 1.000 for classes 'i', 'f', 'e', 'c', and 'a'. This indicates that the model correctly classified a high proportion of instances for each of these classes. The precision of the model, which gauges the proportion of positive identifications that were correct, ranged from 0.892 for class 'h' to a perfect score of 1.000 for classes 'g', 'f', 'd', 'c', and 'b'. This suggests that most instances predicted by the model to belong to each of these classes were indeed correctly classified.

The model's decision-making process can be comprehended by examining the decision tree. The tree generates a hierarchical structure of if-then-else decision rules leading to the classification of an instance. For instance, the initial variable to cause a split is 'Shaft power', segregating instances into those with 'Shaft power' less or equal to 8945.0 and those with 'Shaft power' greater than this value. The decision-making process continues with further splits based on various variables, providing valuable insights into which variables are pivotal in determining the class of an instance.

Finally, the performance of the model on each class is further visualized in the confusion matrix in Table A7. The confusion matrix provides the number of correctly and incorrectly classified instances for each class. For example, out of 39 total instances of class 'g', 35 were correctly classified as 'g', and 4 were misclassified as 'h'. Classes 'i', 'f', 'e', 'd', and 'c' demonstrated perfect classification with all instances correctly classified. Class 'h' had two instances misclassified as 'i', and class 'd' had one instance misclassified as 'e'. Overall, the confusion matrix reaffirms the high accuracy of the model and provides detailed insights into its performance on a class level.

The employed ML models may appear simplistic due to the availability of relevant data, their accuracy, and the straightforward physical laws underlying the modeled data. However, the data modeling process revealed that achieving interpretability often necessitates simplification. One effective approach to simplification is to focus on a single output variable or on those variables that have a dominant influence on it, allowing for a more man-

ageable analysis that can be further decomposed as needed. Complex models, involving highly branched decision trees with numerous variables, can be challenging to comprehend, explain, and validate. This study demonstrates that simplification, achieved through selective focus and stepwise decomposition, is essential for making models comprehensible. It underscores the critical role of collaboration between domain experts to navigate potential pitfalls and ensure that the models remain interpretable, thereby enabling in-depth analyses and validation.

Collaboration between domain experts was critical in selecting the initial set of variables from all the available. The domain expertise clarified the mechanics, confirming that averaging the data from individual cylinders is justifiable when the engine is operating within its optimal performance parameters. This is also affirmed that within the operational temperature range specified by the manufacturer, and provided that scavenge pressure remains within normal bounds, air temperature does not significantly impact fuel consumption. This expert advice informed our methodology for segmenting fuel consumption into discrete categories that adhere to the specific consumption margins outlined in case ship charter party.

Advanced ML tools with user-friendly interfaces now enable any technically proficient person to apply them; however, collaboration with ML practitioners is important to avoid common data analysis pitfalls. In this work, this included data cleaning, which resulted in identification and removal of duplicate data points which would otherwise probably pollute the training and testing data. Visualizing data and making sure the level of data understanding was high enough led to removing inappropriate variables (air temperature and date), which would act as IDs and lead to excellent models' performance on available data but would fail to generalize to new data. Manual feature selection resulted in replacing parameters measured on each of the six cylinders with an average over the cylinders, which simplified further DM steps. Adding weights to training instances and stratified sampling to split the data into the training and test sets was important to balance the importance of days with hundreds vs. days with just a couple of measurements for training and testing. Performing feature selection with multiple methods revealed that the simple correlation-based method (often used as the default) is not appropriate for this dataset because most parameters are highly correlated. The systematic and detail-oriented approach led to models that are simple to understand and are more accurate than models obtained with (semi) automatic methods. Finally, thorough and methodologically sound evaluation resulted in interpretation of the results that puts prediction errors into perspective compared to the actual consumption as well as the baseline prediction model.

In conclusion, all three J48/C4.5 models demonstrate that model complexity, performance, and interpretability are intertwined aspects in ML. Feature selection can enhance interpretability and robustness, but at the potential cost of some accuracy. Conversely, using all variables can boost accuracy but may result in a more complex model that could be challenging to interpret. As such, the choice between these approaches should be guided by the specific needs of the application at hand, balancing the demands for accuracy, complexity, and interpretability.

5. Conclusions

Our study on a VLGC two-stroke engine highlights the importance of interpretability and transparency in machine learning applications within the maritime sector, driven by heightened regulatory demands. We validate the model by incorporating auxiliary criteria, balancing comprehensibility with performance metrics such as accuracy, and providing clear insights into how different input variables affect fuel consumption. This analysis employs two interpretable ML models: linear regression and decision trees, to ensure the interpretability of the models, thereby prioritizing transparency over complexity.

The optimal linear model utilizes shaft power, slip, and exhaust temperature parameters, achieving an RMSE of 23.16 and an MRAE of 14.7%. Decision trees, on the other hand, elucidate the impact of factors such as shaft power, shaft revolutions per minute,

compressor pressure, and indicated mean effective pressure on fuel consumption, achieving accuracy between 96.4% and 97.69%. By visualizing and dissecting the structures of the decision trees, the decision-making process became easily decipherable, allowing the key drivers of fuel consumption to be identified.

Furthermore, this study highlights the importance of close collaboration between domain experts to avoid common ML pitfalls and ensure the interpretability of model behavior. The study is a step toward developing interpretable models for predicting fuel consumption, ensuring that model transparency, accountability, reliability, and safety are maintained.

However, the chosen approach has limitations: the model's simplicity restricts its predictive capabilities and scope. It considers the effects of included variables and assumes their linear impact, offering meaningful insights into relationships but not necessarily implying causality. The second limitation arises from the dataset, which only captures the operations of a VLGC under favorable weather and oceanographic conditions.

Future research will focus on developing interpretable machine learning models to address a broader range of maritime engineering challenges, encompassing various types of marine equipment and operational conditions. Specifically, a collaborative, multidisciplinary approach can be employed to create robust models that enhance decision-making, improve fuel efficiency, extend engine lifespans, and enhance anomaly detection. These advancements are anticipated to make a significant contribution to more sustainable maritime operations.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jmse12101849/s1. Supplementary material Interpretable machine learning VLGC ship fuel consumption.pdf.

Author Contributions: Conceptualization, A.V.; methodology, A.V., S.M.-I. and R.P. software, R.P.; validation, A.V., S.M.-I. and R.P.; formal analysis, A.V., S.M.-I. and R.P.; investigation, A.V.; resources, A.V. and S.M.-I.; data curation, R.P.; writing—original draft preparation, A.V., S.M.-I. and R.P.; writing—review and editing, A.V., S.M.-I. and R.P.; visualization, R.P.; supervision, S.M.-I.; funding acquisition, S.M.-I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available in the Supplementary Materials File.

Acknowledgments: This work has been partially supported by the University of Rijeka under project number uniri-drustv-18-20. A.V. is supported by HORIZON EUROPE Widening participation and spreading excellence INNO2MARE project (grant agreement ID: 101087348).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1. Linear Regression

The formula for multiple linear regression is [21]:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p + \varepsilon \quad \text{(A1)}$$

where: y is the dependent variable (fuel consumption), b_0 is the intercept term, b_i are the coefficients (slopes) associated with each independent variable x_i, and ε represents the error term, accounting for the difference between the observed and predicted values.

The goal of multiple linear regression is to estimate the coefficients bi that minimize the sum of squared differences between the observed and predicted values. Once the coefficients are estimated, the model can predict the dependent variable y for new sets of independent variables xi by plugging them into the formula.

The quality of regression models (i.e., how well a regression model explains the variability in the dependent variable) can be evaluated using multiple metrics [21]. R-squared, often denoted as R^2 is a value between 0 and 1 that quantifies the proportion of

the variance in the dependent variable that can be explained by the independent variables in the model. Value of 1 means a perfect model while value of 0 means a completely useless model. The Root Mean Squared Error (RMSE) is another important performance indicators for regression models. It measures the average of the squared differences between the actual and predicted values, providing a measure of the average deviation between the predicted and actual values of the dependent variable. The RMSE formula is expressed as follows:

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (Actual_i - Predicted_i)^2} \quad \text{(A2)}$$

where n is the number of data points, $Actual_i$ is the actual value of the dependent variable for the i-th data point and $Predicted_i$ is the predicted value of the dependent variable for the i-th data point. RMSE 0 means a perfect model, while a positive value means that model makes errors. RMSE is expressed in the same units as the predicted value and should be interpreted relative to the range of the dependent variable. For example, an RMSE of 10 for a dependent variable with a range of 100 may be more acceptable than an RMSE of 10 for a dependent variable with a range of 20.

Next evaluation metric is mean absolute error (MAE), which measures the average magnitude of the errors and is calculated as the sum of absolute errors divided by the sample size [21].

$$MAE = 1/n \sum_{i=1}^{n} |Actual_i - Predicted_i| \quad \text{(A3)}$$

Like RMSE, MAE is in the same units as the dependent variable and lower MAE values indicate better model performance, as they suggest smaller errors between predicted and actual values. On the contrary, MAE is less sensitive to outliers compared to RMSE because it does not square the prediction errors. Therefore, it provides a more balanced view of the overall model's performance.

Finally, mean relative absolute error (MRAE) is the ratio between the error of the model and the reference, which is the error of a baseline model that always predicts the average value [21]

$$MRAE = 1/n \sum_{i=1}^{n} \left| \frac{Actual_i - Predicted_i}{Actual_i - mean(Actual_i)} \right|. \quad \text{(A4)}$$

MRAE 0 means a perfect model, 1 means a model that is as good/bad as the model that always predicts the mean value while values above 1 mean that the model is useless because it makes larger errors than the baseline model.

Appendix A.2. Decision Tree

Various criteria exist for selecting the best splitting function [24], such as entropy (H), information gain (IG), Gini index, gain ratio, orthogonality measure (ORT), and chi2 (CHAID) method. The complexity of the resulting tree can be reduced through tree pruning.

In this study, the C4.5 [33,34] algorithm with the J48 implementation [37] of tree building was used. The C4.5 algorithm is an improved version of the ID3 algorithm [34] that begins with the original set S as the root node. In each iteration of the algorithm, all unused variables in set S are considered, and the entropy (H) is calculated. Then, the variable with the lowest entropy (H) or the highest information gain (IG) is selected:

$$IG(S, A) = H(S) - \sum_{t \in T} p(t) H(t) = H(S) - H(S|A) \quad \text{(A5)}$$

where $H(S)$ represents the entropy of set S, T are the subsets resulting from dividing set S based on variable A, $p(t)$ is the ratio of the number of elements in t to the number of elements in set S, and $H(t)$ is the entropy of subset t. Information gain (IG) measures the difference in entropy before and after the division of set S by variable A.

Set S is then split into subsets based on the selected variable values A to produce data subsets in the assumed binary space. Some subsets contain positive and negative examples, and the entropy is calculated as follows:

$$\mathrm{H(S)} = -\mathrm{p}_{+}\mathrm{log}_2\mathrm{p}_{+} - \mathrm{p}_{-}\mathrm{log}_2\mathrm{p}_{-} \tag{A6}$$

where $p+$ is the ratio of positive examples in the set S, and $p-$ is the ratio of negative examples in set S. Entropy represents the minimum number of bits required to encode arbitrary members of set S and, in the case of transitioning from binary to c classes, is given by:

$$\mathrm{H(S)} = \sum_{i=1}^{c} -\mathrm{p_i}\mathrm{log}_2\mathrm{p_i}. \tag{A7}$$

The algorithm is repeated on each subset/tree node, considering only the remaining variables that have not been selected for splitting in the nodes above the node to be split. The process is performed recursively until the set in a given subtree is homogeneous, meaning it contains objects that belong to the same category or class. The C4.5 method, unlike ID3, can handle both continuous and discrete variables, handle incomplete data, and solve the problem of overfitting through tree pruning.

Appendix B

The results are reported using standard machine learning evaluation metrics, as defined in numerous textbooks [21,22]: Accuracy, True positive rate (TP Rate), False positive rate (FP Rate), precision, recall, F1-measure, Area under receiver operating characteristic curve (AUROC), Area under precision-recall curves (AUPRC).

Accuracy is the proportion of correctly classified instances in the total set of instances.

True Positive Rate (TP Rate): Also known as recall, is the ratio of the number of correctly classified positive cases (true positives, TP) to the total number of positive values (P):

$$TPR = \frac{TP}{P}. \tag{A8}$$

Total number of positive values (P): This is equal to the sum of correctly classified positives (true positives, TP) and false negatives (FN):

$$\mathrm{P} = \mathrm{TP} + \mathrm{FN}. \tag{A9}$$

False Positive Rate (FP Rate) is equal to the ratio of the number of incorrectly classified positive cases (false positives, FP) to the total number of negative values (N):

$$FPR = \frac{FP}{N}. \tag{A10}$$

Total number of negative values (N): This is equal to the sum of incorrectly classified positives (false positives, FP) and correctly classified negatives (true negatives, TN):

$$N = FP + TN. \tag{A11}$$

Basic evaluation measures are commonly presented in a confusion matrix. The confusion matrix is used to describe the performance of a classifier when the actual value of the data being evaluated is known. The results are often displayed in the form of a two-dimensional confusion matrix (Table A1). The positive/negative labels refer to the predicted outcome of the experiment, while true/false refer to the actual outcome. Positive results correspond to the numbers on the main diagonal, and negative results are in the off-diagonal cells, which ideally have low, ideally zero, values.

Table A1. Confusion matrix.

Predicted	Actual	
	Positive	**Negative**
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

Precision is the proportion of correctly classified cases among the retrieved cases. Recall interface is equivalent to the true positive rate (TPR). Both precision and recall are calculated from the confusion matrix of outcomes as follows:

$$PPV = \frac{TP}{TP + FP}. \tag{A12}$$

$$PTPR = \frac{TP}{TP + FN}. \tag{A13}$$

F1 score is calculated as the harmonic mean of precision (positive predictive value, PPV) and the true positive rate (TPR), also known as recall. The formula for the F1 score is:

$$TF = \frac{2 \times PPV \times TPR}{PPV + TPR}. \tag{A14}$$

F1-measure combines precision and recall into a single metric by calculating their harmonic mean, providing a balance between the two, which is particularly useful when dealing with imbalanced datasets.

The Receiver Operating Characteristics (ROC) curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The area under the ROC curve is commonly used to compare the outcomes of classification models and is denoted as AUROC (Area Under the Receiver Operating Curve). Similarly, the area under the precision-recall curve is referred to as AUPRC (Area Under Precision-Recall Curve) [21].

Kappa (K) statistics, also known as Cohen's Kappa, is a measure used to evaluate the performance of J48/C4.5 decision tree models. Kappa measures the agreement between the classifications made by a model and the actual outcomes, adjusted for agreement that could occur by chance. It is calculated as [21]:

$$\mathrm{K} = \frac{P_o - P_e}{1 - P_e} \tag{A15}$$

where P_o is the observed agreement (i.e., accuracy), and P_e is the expected agreement by chance. A Kappa value of 1 indicates perfect agreement, 0 indicates no better than chance, and negative values suggest worse than random chance.

Appendix B.1. Decision Tree J48/C4.5 Model Performance (RPM/FO)

Evaluation results:
Correctly Classified Instances 695; 68.2711%
Incorrectly Classified Instances 323; 31.7289%
Kappa statistic 0.6439

Table A2. True positive rate (TP Rate), False positive rate (FP Rate), precision, recall, F1-measure, Matthew's correlation coefficient (MCC), Area under receiver operating characteristic curve (AUROC), Area under precision-recall curves (AUPRC) per Class for RPM/FO model.

TP Rate	FP Rate	Precision	Recall	F1-Measure	MCC	AUROC	AUPRC	Class
0.762	0.010	0.914	0.762	0.831	0.814	0.986	0.895	g
0.825	0.053	0.673	0.825	0.742	0.708	0.972	0.763	h
0.369	0.019	0.633	0.369	0.466	0.450	0.939	0.492	i
0.682	0.055	0.593	0.682	0.635	0.590	0.937	0.589	f
0.979	0.192	0.342	0.979	0.507	0.517	0.901	0.341	e
0.177	0.000	1.000	0.177	0.301	0.401	0.900	0.462	d
0.447	0.010	0.875	0.447	0.592	0.588	0.931	0.618	c
0.941	0.006	0.957	0.941	0.949	0.942	0.996	0.946	a
0.957	0.008	0.940	0.957	0.948	0.942	0.996	0.952	b

Table A3. Confusion Matrix.

a	b	c	d	e	f	g	h	i	<-- Classified as
96	29	0	1	0	0	0	0	0	a = g
0	99	1	20	0	0	0	0	0	b = h
9	17	31	27	0	0	0	0	0	c = i
0	0	16	73	18	0	0	0	0	d = f
0	1	0	1	92	0	0	0	0	e = e
0	1	1	0	82	20	9	0	0	f = d
0	0	0	1	77	0	63	0	0	g = c
0	0	0	0	0	0	0	111	7	h = a
0	0	0	0	0	0	0	5	110	i = b

Appendix B.2. Decision Tree J48/C4.5 Model Performance (Power/FO)

Evaluation results:
Correctly Classified Instances 986; 96.8566%
Incorrectly Classified Instances 32; 3.1434%
Kappa statistic 0.9646

Table A4. True positive rate (TP Rate), False positive rate (FP Rate), precision, recall, F1-measure, Matthew's correlation coefficient (MCC), Area under receiver operating characteristic curve (AUROC), Area under precision-recall curves (AUPRC) per Class for Power/FO model.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	AUROC	AUPRC	Class
0.889	0.001	0.991	0.889	0.937	0.931	0.998	0.976	g
0.950	0.016	0.891	0.950	0.919	0.909	0.997	0.969	h
1.000	0.011	0.894	1.000	0.944	0.940	0.997	0.930	i
0.953	0.001	0.990	0.953	0.971	0.968	0.997	0.970	f
0.989	0.000	1.000	0.989	0.995	0.994	0.999	0.995	e
0.973	0.003	0.973	0.973	0.973	0.970	0.998	0.984	d
0.979	0.003	0.979	0.979	0.979	0.975	0.999	0.984	c
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	a
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	b

Table A5. Confusion Matrix.

a	b	c	d	e	f	g	h	i	<-- Classified as
112	14	0	0	0	0	0	0	0	a = g
0	114	6	0	0	0	0	0	0	b = h
0	0	84	0	0	0	0	0	0	c = i
1	0	4	102	0	0	0	0	0	d = f
0	0	0	1	93	0	0	0	0	e = e
0	0	0	0	0	110	3	0	0	f = d
0	0	0	0	0	3	138	0	0	g = c
0	0	0	0	0	0	0	118	0	h = a
0	0	0	0	0	0	0	0	115	i = b

Appendix B.3. Decision Tree J48/C4.5 Model Performance (ALL Variables)

Evaluation results:
Correctly Classified Instances 338; 97.6879%
Incorrectly Classified Instances 8; 2.3121%
Kappa statistic 0.9739

Table A6. True positive rate (TP Rate), False positive rate (FP Rate), precision, recall, F1-measure, Matthew's correlation coefficient (MCC), Area under receiver operating characteristic curve (AUROC), Area under precision-recall curves (AUPRC) per Class for all variables model.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	AUROC	AUPRC	Class
0.897	0.000	1.000	0.897	0.946	0.941	0.994	0.953	g
0.943	0.013	0.892	0.943	0.917	0.907	0.993	0.953	h
1.000	0.006	0.938	1.000	0.968	0.965	1.000	0.994	i
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	f
1.000	0.003	0.963	1.000	0.981	0.980	1.000	0.999	e
0.978	0.000	1.000	0.978	0.989	0.987	1.000	1.000	d
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	c
1.000	0.003	0.975	1.000	0.987	0.986	0.998	0.975	a
0.977	0.000	1.000	0.977	0.989	0.987	0.989	0.980	b

Table A7. Confusion Matrix.

a	b	c	d	e	f	g	h	i	<-- Classified as
35	4	0	0	0	0	0	0	0	a = g
0	33	2	0	0	0	0	0	0	b = h
0	0	30	0	0	0	0	0	0	c = i
0	0	0	37	0	0	0	0	0	d = f
0	0	0	0	26	0	0	0	0	e = e
0	0	0	0	1	45	0	0	0	f = d
0	0	0	0	0	0	50	0	0	g = c
0	0	0	0	0	0	0	39	0	h = a
0	0	0	0	0	0	0	1	43	i = b

Appendix C

Scatterplots and Correlations

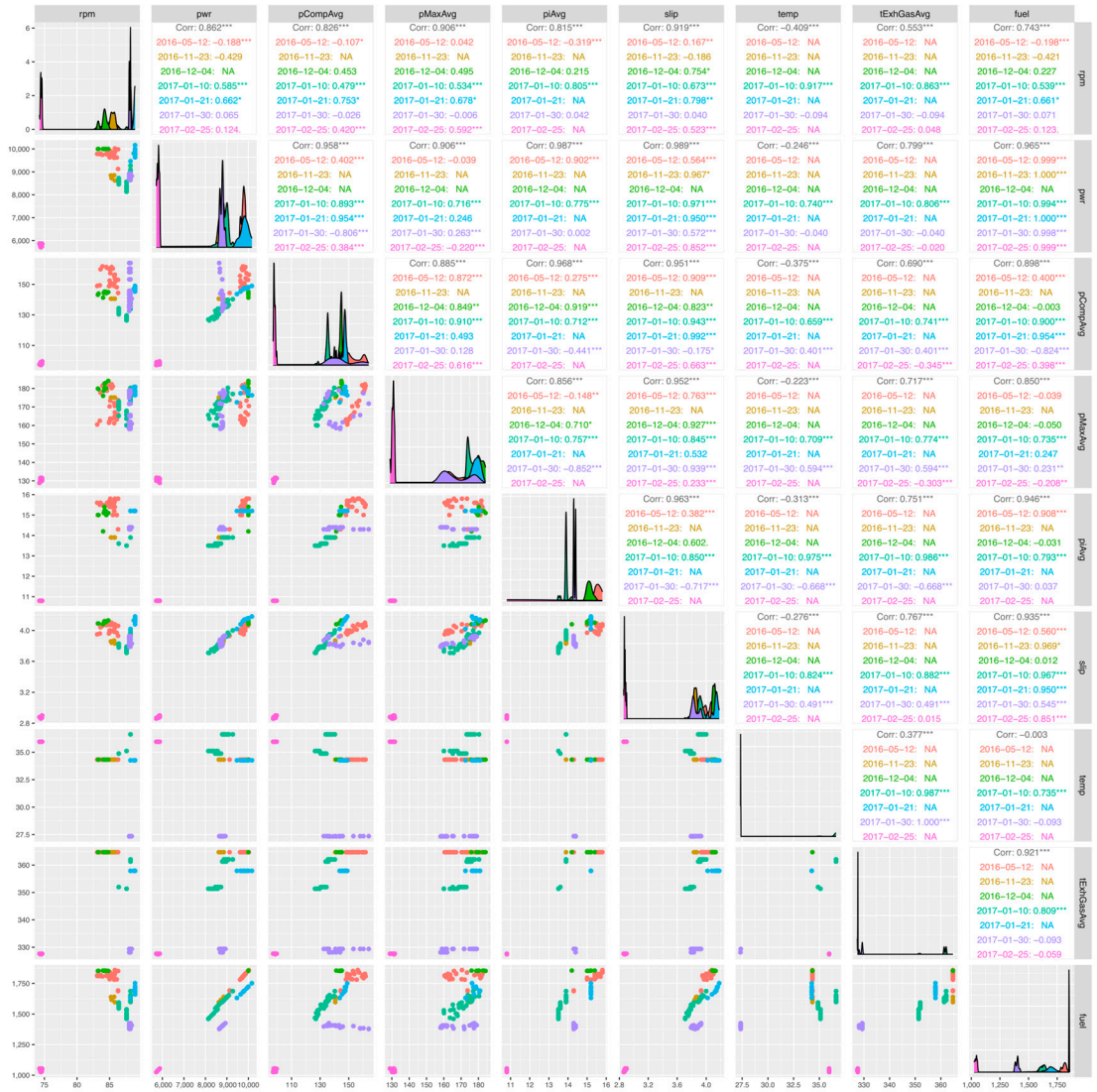

Figure A1. Pairwise parameter scatter plots and Pearson correlations, colors correspond to the dates when data was collected.

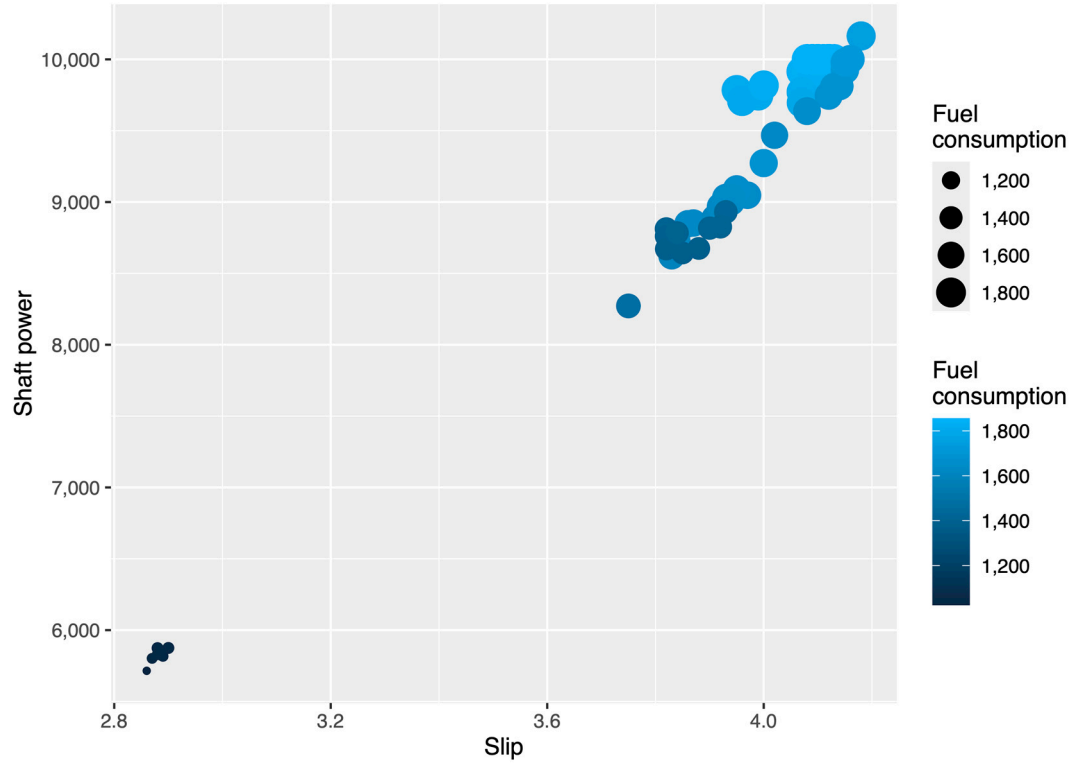

Figure A2. Slip and shaft power vs. fuel consumption.

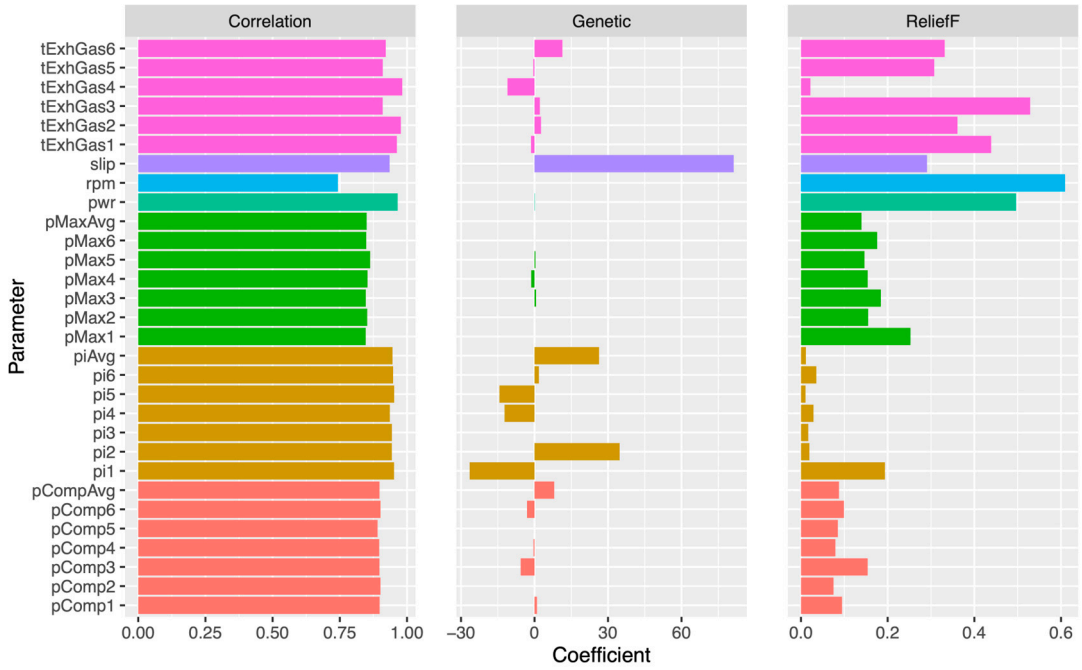

Figure A3. Feature selection on all variables using correlation, genetic and RelieF methods. Figure shows that the simple correlation-based feature selection method fails in case of multiple correlated features (e.g., same parameter measured at each of the 6 engine cylinders) and that data understanding is the key for removing such redundant features (i.e., understanding which features should be removed and why). Furthermore, it demonstrates that advanced feature selection methods such as ReliefF can narrow down the set of useful features much better even when many redundant features are present.

References

1. Jin, W.; Gan, H.; Cong, Y.; Li, G. Performance Optimization and Knock Investigation of Marine Two-Stroke Pre-Mixed Dual-Fuel Engine Based on RSM and MOPSO. *J. Mar. Sci. Eng.* **2022**, *10*, 1409. [CrossRef]
2. Kim, Y.-C.; Kim, K.-S.; Yeon, S.; Lee, Y.-Y.; Kim, G.-D.; Kim, M. Power Prediction Method for Ships Using Data Regression Models. *J. Mar. Sci. Eng.* **2023**, *11*, 1961. [CrossRef]
3. Nielsen, R.E.; Papageorgiou, D.; Nalpantidis, L.; Jensen, B.T.; Blanke, M. Machine learning enhancement of maneuvering prediction for ship Digital Twin using full-scale recordings. *Ocean Eng.* **2022**, *257*, 11579. [CrossRef]
4. Vorkapić, A.; Radonja, R.; Martinčić-Ipšić, S. Predicting Seagoing Ship Energy Efficiency from the Operational Data. *Sensors* **2021**, *21*, 2832. [CrossRef] [PubMed]
5. Vorkapić, A.; Radonja, R.; Martinčić-Ipšić, S. A framework for the application of shipboard energy efficiency monitoring, operational data prediction and reporting. *Pomorstvo* **2021**, *35*, 3–15. [CrossRef]
6. Xie, X.; Sun, B.; Li, X.; Olsson, T.; Maleki, N.; Ahlgren, F. Fuel Consumption Prediction Models Based on Machine Learning and Mathematical Methods. *J. Mar. Sci. Eng.* **2023**, *11*, 738. [CrossRef]
7. Vorkapić, A.; Radonja, R.; Martinčić-Ipšić, S. Machine learning methods in monitoring operating behavior of marine two-stroke diesel engine. *Transport* **2020**, *35*, 474–485. [CrossRef]
8. Coraddu, A.; Oneto, L.; Baldi, F.; Cipollini, F.; Atlar, M.; Savio, S. Data-driven ship digital twin for estimating the speed loss caused by marine fouling. *Ocean Eng.* **2019**, *186*, 106063. [CrossRef]
9. Campos, R.M.; Costa, M.O.; Almeida, F.; Guedes Soares, C. Operational Wave Forecast Selection in the Atlantic Ocean Using Random Forests. *J. Mar. Sci. Eng.* **2021**, *9*, 298. [CrossRef]
10. Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. 2017. Available online: https://arxiv.org/abs/1702.08608 (accessed on 15 April 2024).
11. Hempel, C.; Oppenheim, P. Studies in the logic of explanation. *Philos. Sci.* **1948**, *15*, 135–175. [CrossRef]
12. Bechtel, W.; Abrahamsen, A. Explanation: A mechanist alternative. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 421–441. [CrossRef] [PubMed]
13. Chater, N.; Oaksford, M. Speculations on human causal learning and reasoning. *Inf. Sampl. Adapt. Cog.* **2006**, 210–236.
14. Freitas, A.A. Comprehensible classification models: A position paper. *ACM SIGKDD Explor. Newsl.* **2014**, *15*, 1–10. [CrossRef]
15. Piltaver, R.; Luštrek, M.; Gams, M.; Martinčić-Ipšić, S. What makes classification trees comprehensible? *Expert Syst. Appl.* **2016**, *16*, 333–346. [CrossRef]
16. Wang, T.; Rudin, C.; Velez-Doshi, F.; Liu, Y.; Klampfl, E.; MacNeille, P. A bayesian framework for learning rule sets for interpretable classification. *J. Mach. Learn. Res.* **2017**, *18*, 1–37.
17. Wang, T.; Qihang, L. Hybrid predictive models: When an interpretable model collaborates with a black-box model. *J. Mach. Learn. Res.* **2021**, *22*, 1–38.
18. Goodman, B.; Flaxman, S. European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Mag.* **2017**, *38*, 50–57. [CrossRef]
19. Panigutti, C.; Hamon, R.; Hupont, I.; Fernandez Llorca, D.; Fano Yela, D.; Junklewitz, H.; Gomez, E. The role of explainable AI in the context of the AI Act. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, 12–15 June 2023; pp. 1139–1150. [CrossRef]
20. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
21. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. *An Introduction to Statistical Learning: With Applications in Python*; Springer Nature: Berlin/Heidelberg, Germany, 2023.
22. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining Practical Machine Learning Tools and Techniques*; Elsevier-Todd Green: Cambridge, UK, 2017.
23. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on The practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000; Volume 1, pp. 29–39.
24. Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 1984. [CrossRef]
25. Elshawi, R.; Al-Mallah, M.H.; Sakr, S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 146. [CrossRef]
26. Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; Zhu, J. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In *Natural Language Processing and Chinese Computing: Proceedings of the 8th cCF International Conference, NLPCC 2019, Dunhuang, China, 9–14 October 2019*; Springer: Cham, Switzerland, 2019; p. 11839. [CrossRef]
27. Lundberg, S.M.; Lee, S. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; NISP: Long Beach, CA, USA, 2017; Volume 30, pp. 4765–4774.
28. Kepplinger, D. gaselect: Genetic Algorithm (GA) for Variable Selection from High-Dimensional Data. R Package Version 1.0.21. 2023. Available online: https://CRAN.R-project.org/package=gaselect (accessed on 1 December 2023).
29. Kononenko, I.; Šimec, E.; Robnik-Šikonja, M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Appl. Intell.* **1997**, *7*, 39–55. [CrossRef]

30. Robnik-Šikonja, M.; Savicky, P. CORElearn: Classification, Regression and Feature Evaluation. R Package Version 1.56.0. 2021. Available online: https://CRAN.R-project.org/package=CORElearn (accessed on 1 December 2023).
31. *ISO 11631:1998*; Measurement of Liquid Flow—Methods of Specifying Flowmeter Performance. International Organization for Standardization: Geneva, Switzerland, 1998. Available online: https://www.iso.org/obp/ui/en/#iso:std:iso:11631:ed-1:v1 (accessed on 12 October 2023).
32. *ISO 8217:2017*; Petroleum Products—Fuels (Class F)—Specifications of Marine Fuels. International Organization for Standardization: Geneva, Switzerland, 2017. Available online: https://www.iso.org/standard/64247.html. (accessed on 12 October 2023).
33. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1993.
34. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
35. Wilkinson, G.N.; Rogers, C.E. Symbolic descriptions of factorial models for analysis of variance. *Appl. Stat.* **1973**, *22*, 392–399. [CrossRef]
36. Recursive Partitioning and Regression Trees R Package, rpart. Available online: https://github.com/bethatkinson/rpart (accessed on 1 December 2023).
37. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]