forecasting

MDPI

Article

A Reproducible Workflow for AIS-Based ETA Forecasting: Evaluating the Influence of Data Preprocessing and Machine Learning Model Selection

João Marreiros [1,*], Ana de Jesus Mendes [1,2], Marcela Castro [1,3,4], Maria da Graça Costa [1] and Tiago Pinho [1,2]

[1] Escola Superior de Ciências Empresariais, Polytechnic University of Setúbal, Campus do IPS, Estefanilha, 2910-761 Setúbal, Portugal; ana.mendes@esce.ips.pt (A.d.J.M.); marcela.castro@esce.ips.pt (M.C.); graca.costa@esce.ips.pt (M.d.G.C.); tiago.pinho@esce.ips.pt (T.P.)
[2] UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Universidade NOVA de Lisboa, 2829-516 Caparica, Portugal
[3] NECE—Research Centre for Business Sciences, Department of Management and Economics, Faculty of Human and Social Sciences, Universidade da Beira Interior, 6201-001 Covilha, Portugal
[4] CIEQV—Life Quality Research Center, Complexo Andaluz, Aptd 279, 2001-904 Santarém, Portugal
* Correspondence: joao.marreiros@esce.ips.pt

Highlights

What are the main findings?

- Using AIS data from the coastal receiver network of the Administration of the Ports of Sines and the Algarve (APS)-covering Sines, Portimão and Faro-a corrected, leakage-free workflow with voyage-grouped validation showed that Random Forest Regression achieved the most accurate and consistent Estimated Time of Arrival (ETA) forecasts for container vessels calling at the Port of Sines, among three machine learning models tested (KNN, Random Forest, MLP).
- All three machine learning models substantially outperformed simple operational baselines—including the AIS-reported ETA field itself, which performed worse than simply predicting the average arrival time.

What are the implications of the main findings?

- Reliable AIS-based ETA forecasting depends on the joint quality of data preprocessing, feature engineering, and rigorous voyage-level validation—not on machine learning model choice alone—with direct implications for how future studies should guard against data leakage.
- Anomalous vessel behaviour (e.g., loitering) and limited sample sizes remain key constraints on operational deployment, underscoring the need for dedicated anomaly-aware forecasting components and larger, multi-port datasets in future work.

Check for updates

Academic Editor: Shaolong Sun

Received: 30 June 2026
Revised: 31 July 2026
Accepted: 9 August 2026
Published: 11 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Abstract

Accurate Estimated Time of Arrival (ETA) forecasting is essential for improving operational planning and decision-making in modern ports. While machine learning has significantly enhanced ETA prediction using Automatic Identification System data, the impact of data preprocessing on forecasting performance remains underexplored. This study investigates how AIS data preprocessing and machine learning model selection jointly affect ETA forecasting for short-sea shipping, using the Port of Sines as an empirical case study. A reproducible forecasting workflow was developed, integrating dataset construction, voyage selection, feature engineering, predictive modelling and performance evaluation. Three supervised machine learning algorithms, K-Nearest Neighbors, Random Forest

Regression, and Multilayer Perceptron, were trained and compared under identical experimental conditions. The results show that forecasting performance depends not only on model selection but also on the quality of the modelling dataset. In particular, Random Forest Regression achieved the strongest and most consistent performance. It was found to be invariant to feature scaling, whereas scaling had a small negative effect on K-Nearest Neighbors and increased training variance for the Multilayer Perceptron. Although all three models achieved accurate ETA predictions, they exhibited different strengths regarding predictive performance, computational efficiency, and operational applicability. The proposed workflow contributes to the development of transparent and reproducible ETA forecasting methodologies and provides practical guidance for implementing AIS-based decision-support systems in short-sea port operations.

Keywords: ETA forecasting; Automatic Identification System (AIS); machine learning; data preprocessing; feature engineering; short-sea shipping; Port of Sines; maritime logistics; decision support; predictive modelling

1. Introduction

The digital transformation of maritime transport has increased the demand for reliable forecasting methods to support operational decision-making in ports and shipping. Among the different forecasting tasks required in modern port operations, Estimated Time of Arrival (ETA) prediction plays a central role by enabling more effective berth allocation, terminal planning, resource scheduling, and vessel traffic coordination [1,2]. Accurate ETA forecasts contribute to reducing operational uncertainty, improving port efficiency and supporting the growing adoption of data-driven decision-support systems in maritime logistics.

The widespread deployment of the Automatic Identification System (AIS) has created new opportunities for ETA forecasting by providing continuous information on vessel position, speed, course and navigation status [3,4]. Combined with advances in machine learning, AIS data has enabled the development of predictive models capable of learning complex vessel movement patterns from historical trajectories. As a result, numerous studies have demonstrated that machine learning approaches can substantially improve ETA prediction accuracy compared with traditional estimation methods, particularly for dynamic maritime environments characterised by continuously changing navigational conditions [5–7].

Despite these advances, forecasting performance depends not only on the predictive algorithm but also on the quality of the underlying modelling dataset. AIS records frequently contain missing values, duplicated observations, inconsistent destination descriptions, and heterogeneous data formats that require extensive preprocessing before predictive models can be developed [4,8]. Although these preprocessing procedures are commonly applied, they are often treated as preliminary technical activities rather than being systematically evaluated as factors influencing forecasting performance [5,6]. Consequently, it remains unclear to what extent improvements in ETA forecasting arise from the selection of the machine learning model or from the way AIS data are prepared and transformed into modelling datasets.

This study addresses this gap by investigating how AIS data preprocessing and machine learning model selection jointly influence ETA forecasting performance for short-sea shipping, using the Port of Sines as an empirical case study. Specifically, the study assesses the suitability of AIS data for short-range ETA forecasting, develops a reproducible workflow for dataset construction and predictive modelling, compares the performance of

K-Nearest Neighbors (KNN), Random Forest Regression (RFR) and Multilayer Perceptron (MLP), analyses the influence of preprocessing decisions on forecasting accuracy and discusses the operational implications of ETA forecasting for port decision support.

The study makes three principal contributions. First, it proposes a transparent and reproducible workflow integrating dataset construction, feature engineering, and machine learning model development for AIS-based ETA forecasting. Second, it demonstrates that forecasting performance depends on the combined influence of data preprocessing and model selection rather than on algorithm choice alone. Third, it provides practical evidence supporting the implementation of AIS-based forecasting methods for operational decision support in short-sea port environments. Together, these contributions advance current knowledge on maritime ETA forecasting while providing a methodological framework that can support future forecasting applications using AIS data.

The remainder of this paper is organised to reflect the sequential stages of the proposed forecasting workflow. Section 2 reviews the literature on ETA forecasting, AIS data preprocessing, and machine learning applications in maritime logistics, identifying the research gap addressed in this study. Section 3 presents the research design, followed by Section 4, which describes the dataset construction process, including data preprocessing, voyage selection and feature engineering. Section 5 introduces the selected machine learning models, training procedures and performance evaluation metrics. Section 6 presents experimental results, including the effects of feature scaling, model performance and practical implications for model selection. Section 7 discusses the findings in relation to previous research and examines their implications for ETA forecasting and port operations. Finally, Section 8 summarises the main conclusions, contributions, limitations and directions for future research.

2. Theoretical and Bibliographic Background

2.1. ETA Forecasting in Maritime Logistics

Accurate Estimated Time of Arrival (ETA) forecasting has become an essential capability for modern maritime logistics, supporting operational planning and coordination across ports, shipping companies, and terminal operators. Reliable ETA predictions facilitate berth allocation, vessel traffic management, resource scheduling and hinterland logistics, contributing to improved operational efficiency and reduced waiting times [8]. As ports continue to adopt digital decision-support systems, ETA forecasting has emerged as a key component of data-driven port management.

Traditionally, ETA estimation relied on static voyage plans, historical averages and navigational experience. However, these approaches often fail to capture the dynamic nature of vessel movements, where weather conditions, traffic congestion, operational delays and changes in vessel speed continuously influence arrival times [9]. The widespread availability of Automatic Identification System (AIS) data has transformed this scenario by enabling continuous monitoring of vessel trajectories and supporting the development of real-time forecasting models.

Recent advances in machine learning have further improved ETA forecasting by allowing predictive models to learn complex relationships from historical vessel trajectories. Unlike conventional statistical approaches, machine learning algorithms can capture nonlinear interactions among navigational variables and continuously adapt to large volumes of AIS observations, resulting in more accurate and robust ETA predictions [5,8].

Despite these advances, ETA forecasting remains a challenging problem. Prediction accuracy depends not only on the forecasting algorithm but also on the quality, consistency and representativeness of the underlying navigational data. AIS datasets frequently contain missing values, duplicated records, inconsistent destination descriptions and heterogeneous data formats, all of which may compromise model performance if not adequately addressed

during data preparation [1,10]. Consequently, recent research has increasingly recognised that improving ETA forecasting requires not only advances in predictive algorithms but also robust methodologies for constructing reliable modelling datasets.

2.2. AIS Data and Data Preprocessing

The Automatic Identification System (AIS) has become the primary data source for vessel trajectory analysis and ETA forecasting because it provides continuous information on vessel identity, position, speed, course, and navigational status (Table 1). Originally developed to improve maritime safety and collision avoidance, AIS has evolved into a valuable source of operational data supporting a wide range of maritime analytics, including vessel traffic monitoring, route optimisation and arrival time prediction [8].

Table 1. Data obtained from the AIS system. Source: Park et al. [3].

Static Information	Dynamic Information	Voyage-Related Information
MMSI	Position timestamp (Coordinated Universal Time)	Ship's draught
Call sign and name	Course over ground (COG)	Hazardous cargo (type)
IMO number	Speed over ground (SOG)	Destination and ETA
Length and beam	Heading	Route plan (waypoints)
Type of ship	Navigational status Rate of Turn	
Location of the position-fixing antenna	Ship's position with accuracy indication and integrity status	

The availability of large-scale AIS datasets has enabled the application of machine learning techniques capable of extracting complex navigation patterns from historical vessel trajectories. Compared with traditional forecasting approaches, AIS-based predictive models benefit from high-frequency observations that capture the dynamic behaviour of vessels throughout their voyages, providing richer information for estimating remaining travel time [5,6].

Despite these advantages, AIS data presents important quality challenges that directly affect forecasting performance. Missing values, duplicated observations, inconsistent destination records, irregular transmission intervals, and heterogeneous data formats are frequently reported in operational AIS datasets [1,10]. Since many AIS attributes are manually entered or collected under varying operational conditions, raw datasets often require extensive preprocessing before they can be used for predictive modelling.

Recent studies have emphasised that data preprocessing should not be regarded merely as a preliminary technical task but as an essential stage of the forecasting workflow. Procedures such as data cleaning, trajectory reconstruction, voyage identification, destination standardisation and feature engineering improve data consistency and enable machine learning models to learn more representative navigation patterns [5,8,10]. These procedures also contribute to the reproducibility of forecasting experiments by establishing transparent and consistent data preparation protocols.

Although previous studies recognise the importance of data quality, comparatively little attention has been devoted to systematically evaluating how preprocessing decisions influence ETA forecasting performance. Understanding the contribution of data preparation to predictive accuracy remains an important research challenge, particularly for short-sea shipping applications where forecasting decisions rely almost exclusively on AIS-derived information.

2.3. Machine Learning for ETA Forecasting

The increasing availability of AIS data has accelerated the adoption of machine learning techniques for ETA forecasting, enabling the development of predictive models capable

of learning complex relationships from historical vessel trajectories. Unlike conventional statistical approaches, machine learning algorithms can capture nonlinear interactions among navigational variables, making them particularly suitable for modelling the dynamic behaviour of maritime transport systems [8].

A wide variety of supervised learning algorithms have been investigated for ETA forecasting, including k-Nearest Neighbors, Random Forest, Support Vector Regression, Gradient Boosting, and Artificial Neural Networks. Comparative studies generally report that no single algorithm consistently outperforms all others, as predictive performance depends on factors such as voyage characteristics, feature selection, data quality, and the operational context in which the models are applied [8,11,12].

Among these approaches, Random Forest has emerged as one of the most frequently adopted algorithms because of its ability to model nonlinear relationships while maintaining robustness to noisy AIS observations [8]. K-Nearest Neighbors has also demonstrated competitive performance when applied to carefully selected navigational variables, particularly in situations where vessel trajectories exhibit similar movement patterns [5]. More recently, Artificial Neural Networks, including Multilayer Perceptron, have shown promising results owing to their capacity to learn complex interactions among multiple spatial and temporal variables derived from AIS data [6].

Although these studies have significantly advanced ETA forecasting, most have concentrated on comparing predictive algorithms or improving model accuracy. Comparatively little attention has been devoted to understanding how the quality of the modelling dataset, feature engineering and preprocessing decisions influence forecasting performance. Consequently, it remains difficult to determine whether improvements in ETA prediction arise primarily from the selected machine learning algorithm or from the way AIS data are prepared before model development. This distinction is particularly relevant for short-sea shipping, where forecasting models often rely almost exclusively on AIS-derived information.

2.4. Research Gap and Research Objectives

The literature demonstrates that significant progress has been achieved in AIS-based ETA forecasting through the application of increasingly sophisticated machine learning techniques. Existing studies have shown that accurate ETA predictions can be obtained using a variety of supervised learning algorithms and that AIS data provide a valuable foundation for developing data-driven forecasting models [5,6,8]. Nevertheless, the emphasis of previous research has been placed primarily on improving predictive algorithms and comparing their forecasting accuracy.

Comparatively less attention has been devoted to understanding how data preprocessing and feature engineering influence the predictive performance of machine learning models. Although data cleaning, trajectory reconstruction and feature generation are routinely performed before model development, these stages are often described only briefly and are seldom evaluated as factors that may directly affect forecasting performance [1,10]. So, the relative contribution of dataset preparation and model selection to ETA forecasting accuracy remains insufficiently explored.

This limitation is particularly relevant for short-sea shipping, where ETA forecasting frequently depends almost exclusively on AIS-derived information. In these operational contexts, improvements in data quality and feature engineering may have an impact comparable to or greater than that of selecting a specific machine learning algorithm. A better understanding of these interactions is therefore essential for developing forecasting models that are both accurate and operationally reliable.

To address this gap, the present study investigates how AIS data preprocessing and machine learning model selection jointly influence ETA forecasting performance for short-sea

shipping, using the Port of Sines as an empirical case study. Specifically, the study: (i) assesses the suitability of AIS data for short-range ETA forecasting; (ii) develops a reproducible workflow for dataset construction and predictive modelling; (iii) compares the predictive performance of K-Nearest Neighbours, Random Forest Regression and Multilayer Perceptron models; (iv) analyses the influence of preprocessing decisions on forecasting accuracy; and (v) discusses the operational implications of ETA forecasting for port decision support.

3. Research Design

This study employed an empirical research design to investigate how AIS data preprocessing and machine learning model selection affect ETA forecasting performance in short-sea shipping. The Port of Sines, Portugal, was selected as the case study because of its strategic role within the European maritime transport network and the availability of large-scale AIS data describing real vessel movements under operational conditions.

The research design comprised four sequential stages. First, raw AIS data were collected and transformed into a consistent modelling dataset through a structured workflow that included data preprocessing, voyage selection and feature engineering. Second, a set of predictor variables was defined based on the literature and the operational characteristics of vessel navigation. Third, three supervised machine learning algorithms representing different learning paradigms—K-Nearest Neighbors (KNN), Random Forest Regression (RFR), and Multilayer Perceptron (MLP)—were trained on the same dataset under identical experimental conditions. Finally, the predictive performance of the models was evaluated using complementary accuracy, generalisation and computational performance metrics.

Figure 1 summarises the methodological framework adopted in this study. The workflow emphasises the sequential relationship between dataset construction, predictive modelling and model evaluation, ensuring that each stage contributes systematically to the development of a reproducible ETA forecasting process.

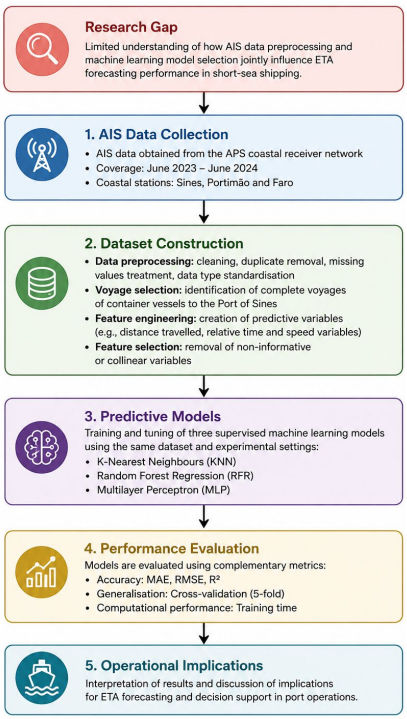


Figure 1. Research design adopted for ETA forecasting using AIS data. Numbers 1–5 indicate the sequential stages of the methodological workflow, from AIS data collection through to the discussion of operational implications; the "Research Gap" box identifies the motivating problem addressed by the study.

The proposed research design directly addresses the gap identified in the literature by evaluating the combined influence of data preprocessing and machine learning model selection within a single experimental framework. Rather than focusing exclusively on algorithm comparison, the study examines how decisions made during dataset construction affect forecast performance, thereby providing a more comprehensive understanding of the factors influencing ETA prediction in short-sea shipping.

4. Dataset Construction

4.1. Data Source

The empirical stage of this research was developed using Automatic Identification System (AIS) data collected by the coastal receiver network operated by the Administration of the Ports of Sines and the Algarve (APS). The dataset covers one year, from June 2023 to June 2024, and includes vessel movements recorded by the AIS stations located in Sines, Portimão and Faro. Incorporating data from the three coastal receivers extended the spatial coverage of vessel trajectories, particularly for ships approaching the Port of Sines from the Mediterranean, thereby increasing the amount of navigational information available for ETA forecasting.

The Port of Sines was selected as the case study because of its strategic importance within the European maritime network and the operational complexity of managing multiple terminals, concessionaires, and high volumes of international traffic. As illustrated in Figure 2, the port combines extensive maritime infrastructure with direct connections to major international shipping routes, making it an appropriate environment for evaluating ETA forecasting models under real operational conditions.



Figure 2. Aerial view of the Port of Sines and its surroundings.

The original dataset consisted of 95 AIS files, totalling approximately 6.8 GB compressed, 45.3 GB after decompression, and 46.46 GB after conversion to a single CSV dataset for processing. Overall, the database contained 301,893,822 AIS records distributed across 17 variables. Each record represents a timestamp transmitted by a vessel while navigating within the reception range of one of the APS coastal stations. Consequently, multiple records describe the complete trajectory of a vessel during a voyage rather than the trajectories of individual ships.

An initial inspection of the raw dataset revealed several characteristics requiring preprocessing before model development. As summarised in Figure 3, the data included heterogeneous variable types, missing values, duplicated observations, manually entered destination fields with inconsistent descriptions, and temporal variables stored as text rather than datetime objects. In addition, the dataset included multiple vessel categories, although only container vessels were relevant to this study's objectives.

```
dataset.shape
{301893822} {17}
```

```
# Checking nulls again
null_summary = dfToClean.isnull().sum()
print(null_summary)
```

```
UTCDate            0
long              16
lat               16
MMSI               0
Imo             2261
tipo_cod           0
ETA            38886
Destination    51923
SOGKnots          74
COGDegrees      4203
dtype: int64
```

```
dataset.dtypes
```

```
UTCDate          object
long            float64
lat             float64
SourceName       object
MMSI              int64
Imo             float64
Name             object
tipo_cod          int64
tipo_desc        object
ETA              object
Destination      object
SOGKnots        float64
COGDegrees      float64
calado          float64
largura         float64
comprimento     float64
Simulated          bool
dtype: object
```

Figure 3. Initial characteristics of the AIS dataset, including dimensions (dataset.shape), missing values (in almost all features), and data types (object, float64, int64 and boolean).

Since ETA forecasting performance depends directly on the quality and consistency of the input data, these characteristics required a structured preprocessing workflow before feature engineering and model training. The following section describes the procedures adopted to transform the raw AIS records into a consistent and reproducible modelling dataset [1,5,10].

4.2. Data Preprocessing

To ensure that the AIS data were suitable for predictive modelling, a structured preprocessing workflow was implemented to improve data quality while preserving the temporal and spatial characteristics of vessel trajectories. The adopted procedure focused on four main stages: variable selection and standardisation, voyage identification, missing-value treatment, and duplicate removal. This workflow was designed to maximise data consistency and establish a reproducible dataset for ETA forecasting [1,5,10].

Initially, variables unrelated to ETA prediction were removed to reduce computational complexity while retaining all information required for trajectory reconstruction and model development. The original ETA field was preserved to enable future comparisons between reported and predicted arrival times, although it was not used as an input variable during model training. Since the objective of this study was to forecast the arrival time of container vessels calling at Terminal XXI, only vessels classified as AIS type code 70 were retained for subsequent processing.

Before trajectory reconstruction, temporal variables were converted to datetime format to support the calculation of travel times and other derived attributes. Destination records also required extensive harmonisation because this field is manually entered by vessel crews and contains numerous spelling variations referring to the Port of Sines. Following consultation with APS specialists, 35 different destination descriptions were identified

and standardised into a single destination code ("PTSINES"), ensuring consistent voyage identification while preserving records with temporarily incomplete information.

To facilitate the reconstruction of complete vessel trajectories, each voyage was assigned a unique identifier (ShipVoyage) created by combining the vessel MMSI with a sequential voyage number. Consecutive voyages from the same vessel were distinguished whenever the interval between successive AIS messages exceeded six hours, indicating the completion of one voyage and the beginning of another. As illustrated in Figure 4, this procedure enabled all AIS messages associated with the same voyage to be consistently grouped across the remaining preprocessing stages. Figure 4 also shows the final naming convention used after variable standardisation.

	MMSI	Destination	BaseDateTime	ShipVoyage
0	209550000.0	PTSines	2023-09-25 12:49:00	209550000_0
1	209550000.0	PTSines	2023-09-25 12:49:02	209550000_0
2	209550000.0	PTSines	2023-09-25 12:49:14	209550000_0
3	209550000.0	PTSines	2023-09-25 12:49:14	209550000_0
4	209550000.0	PTSines	2023-09-25 12:49:20	209550000_0

Figure 4. Construction of the unique ShipVoyage identifier used to reconstruct complete vessel trajectories and use of standardised PTSines name for destination.

Missing values were subsequently treated according to each variable's characteristics rather than applying a single imputation strategy.

Missing destination information was completed whenever consistent values were available within the same ShipVoyage, while missing IMO identifiers were recovered from other observations belonging to the corresponding vessel trajectory. Missing ETA values were estimated from neighbouring records within the same voyage to preserve temporal continuity. In contrast, observations that lacked both positional and navigational information were removed because they could not contribute to reliable trajectory reconstruction. Following the missing-value treatment, duplicated AIS messages generated by simultaneous reception at multiple coastal stations were identified using a combination of vessel identifier, timestamp, and geographical position, ensuring that only genuine duplicates were removed.

The resulting dataset after preprocessing is summarised in Figure 5, which demonstrates the substantial improvement in data completeness and consistency achieved before feature engineering and model development.

UTCDate	0
long	16
lat	16
MMSI	0
Imo	2261
tipo_cod	0
ETA	38886
Destination	51923
SOGKnots	74
COGDegrees	4203
dtype: int64	

BaseDateTime	0
LON	16
LAT	16
MMSI	0
Imo	2261
Type	0
ETA	38886
Destination	51885
SOG	74
COG	4203
ShipVoyage	0
dtype: int64	

Figure 5. Dataset characteristics after preprocessing, including remaining missing values.

The preprocessing workflow resulted in 91 complete and uniquely identified voyages comprising 381,529 AIS records. By systematically addressing inconsistencies prior to feature engineering, the resulting dataset provides a reliable basis for evaluating how data

preparation affects ETA forecasting performance. It ensures that the subsequent machine learning experiments are fully reproducible.

4.3. Voyage Selection

Following data preprocessing, a set of selection criteria was applied to ensure that only complete and operationally representative voyages were included in the modelling dataset. These criteria were established to maximise trajectory consistency while avoiding observations that could introduce bias into ETA forecasting or reduce the predictive capability of the machine learning models.

The first selection criterion was based on the operational boundary of the Port of Sines. Given the port's maritime jurisdiction, a circular area with a 7 km radius centered on the port was designated as the vessel arrival zone. The first AIS record registered within this area was considered the Actual Time of Arrival (ATA), while all subsequent observations were excluded because they no longer contributed to ETA prediction. This approach ensured that each voyage terminated at a consistent operational reference point.

The second criterion established a minimum voyage length. Only trajectories starting at least 50 km from the 7 km arrival boundary were retained for analysis. This threshold ensured that each voyage contained sufficient spatial and temporal information to support meaningful model training and performance evaluation, while excluding very short approaches with limited predictive value.

Examples of the selected trajectories are presented in Figure 6, illustrating complete vessel movements from the initial observation point to the defined arrival boundary. These trajectories provided the temporal continuity required for feature engineering and subsequent ETA prediction.

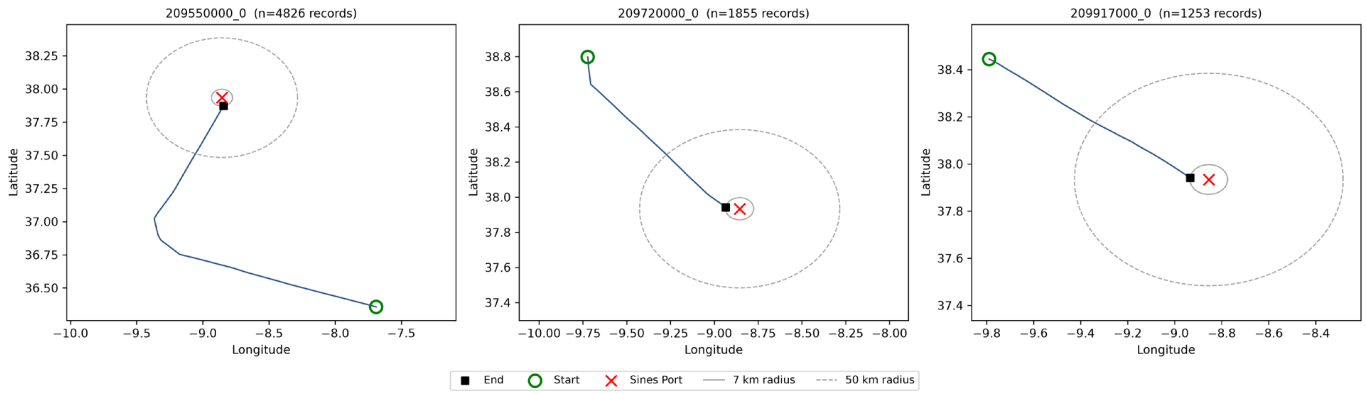


Figure 6. Example of complete voyages retained for model development.

Conversely, voyages that did not reach the arrival boundary or originated within the minimum distance threshold were excluded from the modelling dataset. As illustrated in Figure 7, these trajectories were considered unsuitable because they did not provide sufficient navigational information for reliable ETA estimation.

Finally, five voyages presenting pronounced loitering behaviour (illustrated for one representative case in Figure 8) were removed from the dataset. Such trajectories are characterised by prolonged waiting periods or irregular manoeuvring patterns that are not representative of normal navigation and may adversely affect ETA prediction models. Their exclusion is consistent with previous studies on vessel trajectory prediction that recommend removing anomalous navigation patterns before model development [7].

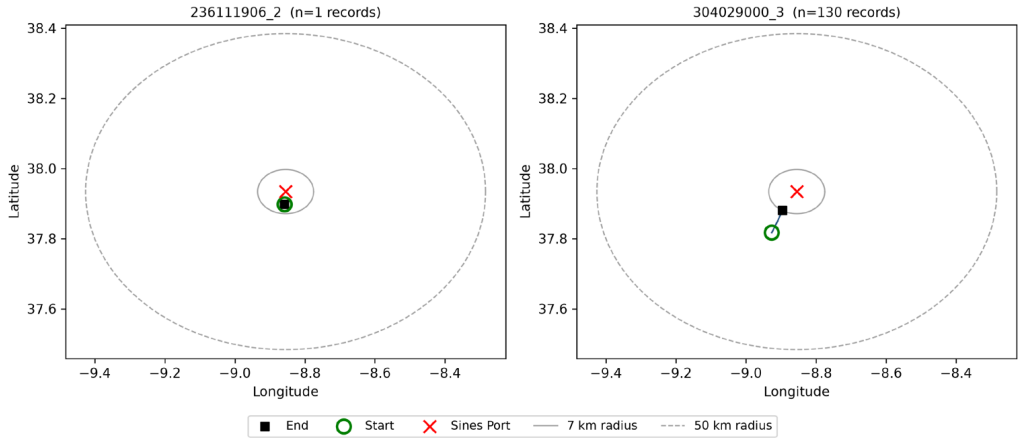


Figure 7. Examples of excluded voyages that did not satisfy the selection criteria.

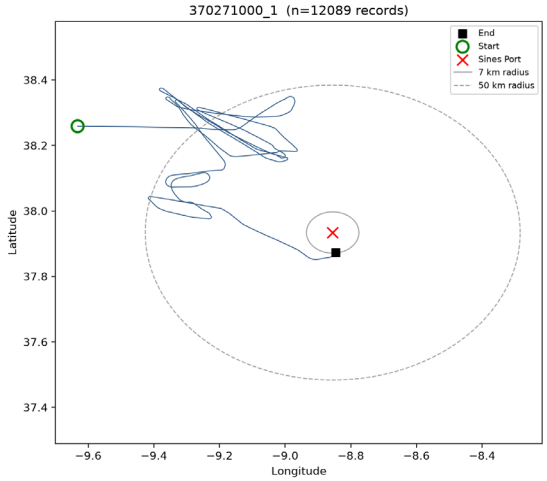


Figure 8. Loitering behaviour of one voyage that could not be included in the model.

After applying these selection criteria, the resulting dataset consisted exclusively of complete and operationally representative voyages, providing a consistent basis for feature engineering and predictive modelling while minimising potential sources of forecasting bias. In total, 16 voyages were excluded during this stage: 11 did not satisfy the arrival-boundary or minimum-length criteria, and 5 exhibited anomalous (loitering-type) behaviour, yielding a final modelling dataset of 75 voyages comprising 304,425 AIS records.

4.4. Feature Engineering

Following voyage selection, additional predictive variables were derived from the processed AIS records to improve the machine learning models' ability to estimate vessel arrival times. Feature engineering focused on extracting temporal and navigational information that could enhance ETA prediction while preserving the operational characteristics of each voyage. This stage transformed the cleaned AIS dataset into a modelling dataset specifically designed for supervised learning [5,8,10].

The first derived variable was the Remaining Travel Distance (RTD), calculated as the great-circle (haversine) distance between the vessel's current position and the fixed coordinates of the Port of Sines. This formulation depends only on information available at the moment of prediction. It does not require knowledge of the vessel's future trajectory, making it directly applicable to real-time operational deployment.

The second derived variable was the Remaining Travel Time (RTT), computed as the difference between the timestamp of each AIS message and the Actual Time of Arrival

(ATA). RTT represents the remaining time until arrival for every observation within a voyage and was adopted as the target variable for model development.

Additional temporal variables describing vessel movements were also derived from AIS timestamps, enabling each observation to capture a voyage's progression over time. Combined with the navigational variables retained during preprocessing, these engineered features provide complementary information describing vessel position, movement, and the remaining distance to the destination.

The final modelling dataset consisted of the original AIS variables retained after pre-processing together with the engineered features generated during this stage. These variables formed the input data for the machine learning models described in the following section, allowing a systematic evaluation of how preprocessing decisions and feature engineering influence ETA forecasting performance. Figure 9 summarises the complete dataset construction workflow adopted in this study, illustrating how raw AIS data were systematically transformed into a modelling dataset suitable for reproducible machine learning experiments. It provides the methodological foundation for evaluating the influence of both data preparation and model selection on ETA forecasting performance.

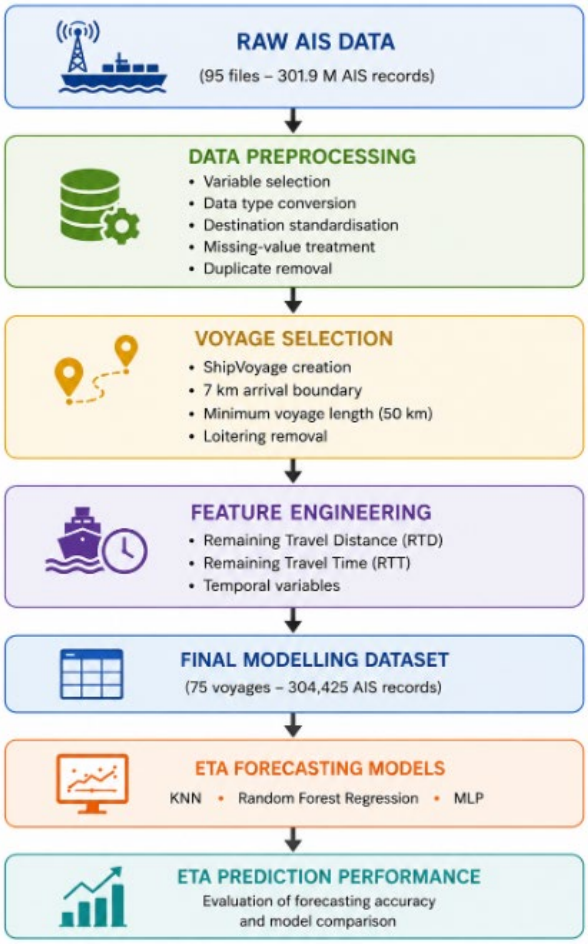


Figure 9. Workflow adopted to construct the final modelling dataset (from raw AIS data to ETA forecasting). Icons are used for illustrative purposes only and correspond to the labelled stage in each box.

5. Predictive Models

5.1. Feature Selection

Feature selection plays a fundamental role in machine learning by identifying the variables that provide the greatest predictive value while reducing redundancy and un-

necessary model complexity. Selecting an appropriate set of features improves model performance, facilitates interpretation, and contributes to more robust and computationally efficient forecasting models, even having operational impacts [13].

Previous studies have shown that ETA forecasting can be achieved using a relatively small set of AIS-derived variables, provided that these variables adequately describe vessel position, movement, and voyage progression. Kolley et al. [5] compared different feature sets and demonstrated that a reduced set of non-standardised navigational variables produced superior predictive performance. Similarly, Valero et al. [8] identified vessel position, distance to destination and speed-related variables as the most influential predictors of ETA, while concluding that weather-related variables provided little additional benefit for short-distance voyages. Magalhães [6] further demonstrated that the Remaining Travel Distance (RTD), combined with Speed over Ground (SOG), substantially improved the estimation of the Remaining Travel Time (RTT), highlighting the importance of engineered variables derived from AIS data.

Based on these findings, the feature selection adopted in this study combined the navigational information available from AIS messages with the engineered variables developed during the dataset construction stage. The selected predictors describe vessel position, movement, dimensions, and voyage progression while avoiding redundant attributes that provide limited additional predictive information.

Table 2 summarises the final set of variables used for model development. The predictor variables include the original AIS attributes retained after preprocessing together with the engineered Remaining Travel Distance (RTD). The Remaining Travel Time (RTT) was adopted as the target variable for supervised learning.

Table 2. Variables used for ETA forecasting model development.

Variable	Source	Description	Role
Latitude	AIS	Vessel latitude at each AIS transmission	Predictor
Longitude	AIS	Vessel longitude at each AIS transmission	Predictor
Speed Over Ground (SOG)	AIS	Vessel speed relative to the ground	Predictor
Course Over Ground (COG)	AIS	Vessel direction of movement over the ground	Predictor
Heading	AIS	Vessel heading measured from true north	Predictor
Navigational Status	AIS	Operational status reported by the vessel	Predictor
Length	AIS	Vessel overall length	Predictor
Width (Beam)	AIS	Vessel overall width	Predictor
Draught	AIS	Reported vessel draught	Predictor
Destination	AIS (standardised)	Standardised destination port (PTSINES)	Predictor
Remaining Travel Distance (RTD)	Engineered	Remaining sailing distance between the vessel's current position and the predefined arrival point	Predictor
Remaining Travel Time (RTT)	Engineered	Remaining time until arrival, used as the response variable	Target

To make explicit which variables are available at the moment of prediction (as opposed to those used only to construct the target label), Table 3 lists the prediction-time availability of every variable used in this study.

The predictor variables combine original AIS attributes with engineered variables derived during dataset construction. RTT was defined as the target variable to be estimated by the machine learning models.

The resulting feature set combines spatial, temporal, and operational information describing vessel behaviour throughout each voyage while keeping the number of input variables manageable. This configuration provides the basis for evaluating how machine learning model selection influences ETA forecasting performance, which is the primary objective of this study.

Table 3. Prediction-time availability of variables used for ETA forecasting.

Variable	Known at Prediction Time?	Notes
Latitude	Yes	Instantaneous AIS position
Longitude	Yes	Instantaneous AIS position
Speed over Ground (SOG)	Yes	Instantaneous AIS attribute
Course over Ground (COG)	Yes	Instantaneous AIS attribute
Heading	Yes	Instantaneous AIS attribute
Navigational Status	Yes	Instantaneous AIS attribute
Length	Yes	Static vessel attribute
Width (Beam)	Yes	Static vessel attribute
Draught	Yes	Reported AIS attribute
Destination (standardised)	Yes	Reported/standardised AIS field
Remaining Travel Distance (RTD)	Yes	Great-circle distance from the current position to the fixed port coordinate; requires no future information
Actual Time of Arrival (ATA)	No	Known only once the voyage is complete; used solely to construct the RTT label, never as a predictor
Remaining Travel Time (RTT)	No–target	Target variable, computed retrospectively from ATA; never available at prediction time

5.2. Model Selection

Selecting an appropriate machine learning algorithm is a critical step in predictive modelling because no single model consistently provides the best performance across all forecasting problems. According to the Algorithm Selection Problem [11] and the No Free Lunch theorem [12], model selection should be guided by the characteristics of the available data and the objectives of the prediction task rather than by the overall popularity or average performance of a particular algorithm. Since the objective of this study is to predict a continuous variable, Remaining Travel Time (RTT), the problem was formulated as a supervised regression task [14].

The literature on ETA forecasting using AIS data reports a wide range of regression algorithms, including Random Forest, Gradient Boosting, Support Vector Regression, k-Nearest Neighbours (KNN), Artificial Neural Networks and Multilayer Perceptron (MLP), among others. A recent review conducted by Valero et al. [8] identified Random Forest as the most frequently applied algorithm for ETA prediction, while Kolley et al. [5] and Magalhães [6] successfully applied KNN and MLP to estimate vessel arrival times using AIS-derived variables.

Based on these findings, three supervised regression models were selected for empirical evaluation: K-Nearest Neighbors (KNN), Random Forest Regression (RFR), and Multilayer Perceptron (MLP). These algorithms were chosen because they represent different modelling approaches, ranging from instance-based learning to ensemble learning and artificial neural networks, allowing the influence of model selection on ETA forecasting performance to be systematically evaluated (Table 4).

Table 4. Machine learning models selected for ETA forecasting.

Model	Learning Approach	Main Strengths	Potential Limitations
KNN	Instance-based learning	Simple implementation, effective for local similarity patterns	Computational cost increases with dataset size
RFR	Ensemble learning	Robust to noise, captures nonlinear relationships, and reduces overfitting	Lower interpretability than linear models
MLP	Artificial neural network	Learns complex nonlinear relationships and high-dimensional patterns	Requires parameter tuning and longer training times

Together, these models provide a representative set of machine learning techniques widely adopted in ETA forecasting studies and enable a comprehensive comparison of how

different modelling strategies influence forecasting accuracy under the same dataset and preprocessing conditions.

5.2.1. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric supervised learning algorithm that predicts the value of a new observation based on the values of the most similar instances in the training dataset. Similarity is typically determined using a distance metric, allowing the algorithm to estimate the target value from neighbouring observations with comparable characteristics. Because KNN makes no assumptions about the underlying data distribution, it is well suited to modelling nonlinear relationships commonly found in AIS trajectory data.

Previous studies have successfully applied KNN to ETA forecasting using AIS data. Kolley et al. [5] demonstrated that KNN achieved competitive predictive performance when combined with a carefully selected set of navigational variables, while Magalhães [6] reported satisfactory results using engineered features such as Remaining Travel Distance (RTD) and Speed over Ground (SOG). These findings indicate that KNN provides a reliable baseline for evaluating ETA prediction performance, particularly in short sea shipping scenarios where neighbouring vessel trajectories frequently exhibit similar navigational patterns.

In this study, KNN was selected as a representative instance-based learning algorithm to evaluate its ability to estimate Remaining Travel Time (RTT) from AIS-derived variables and to compare its performance with ensemble learning and neural network approaches.

5.2.2. Random Forest Regression (RFR)

Random Forest Regression (RFR) is an ensemble learning algorithm that combines the predictions of multiple decision trees to improve predictive accuracy while reducing the risk of overfitting. Each tree is constructed from a bootstrap sample of the training data, and random subsets of predictor variables are considered during the splitting process, increasing model robustness and generalisation capability.

Among the machine learning techniques applied to ETA forecasting, Random Forest has consistently demonstrated strong predictive performance. Valero et al. [8] identified Random Forest as the most frequently adopted algorithm in recent ETA forecasting studies, highlighting its ability to capture complex nonlinear relationships between navigational variables and vessel arrival times. Its robustness to noisy data and reduced sensitivity to outliers make it particularly suitable for AIS datasets, which often contain measurement variability and heterogeneous vessel behaviours.

Given these characteristics, Random Forest Regression was selected as the ensemble learning approach in this study, enabling a comparison of its forecasting performance with both a distance-based algorithm (KNN) and a neural network model (MLP).

5.2.3. Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a feed-forward artificial neural network capable of learning complex nonlinear relationships through interconnected layers of artificial neurons. During training, the network iteratively adjusts its internal weights using backpropagation, enabling it to model intricate interactions among predictor variables without requiring explicit assumptions regarding their functional relationships.

Neural network models have attracted increasing attention in ETA forecasting because of their capacity to capture complex navigation patterns from AIS data. Magalhães [6] successfully applied MLP to Remaining Travel Time prediction, while recent studies have recognised artificial neural networks as effective alternatives for modelling highly nonlinear maritime trajectory data [8]. Their flexibility makes them particularly well-suited to forecasting problems involving multiple spatial and temporal variables, where conventional regression techniques may struggle to capture the underlying relationships.

In the present study, MLP was selected as the neural network approach to assess its capability to model complex interactions among AIS-derived variables and engineered features. Comparing its performance with KNN and Random Forest Regression provides a comprehensive evaluation of three distinct machine learning paradigms for ETA forecasting under identical preprocessing and training conditions.

5.3. Model Training

To ensure a fair and reproducible comparison, all three machine learning models were trained on the same preprocessed dataset, with identical predictor variables and the same target variable (Remaining Travel Time, RTT). This experimental design ensured that any differences in forecasting performance could be attributed to the learning algorithms rather than to variations in data preparation or feature selection.

Before model training, exploratory data analysis (EDA) was conducted to examine feature distributions, identify multicollinearity, and assess the relationships between predictor variables and the target variable. Correlation analysis and the Variance Inflation Factor (VIF) were used to support feature selection and verify that the final modelling dataset was suitable for supervised learning.

Voyages were split at the voyage level (grouped by the ShipVoyage identifier) into an 80/20 training/test partition, consistent with the recommendation for large datasets proposed by Gholamy et al. [15]. To preserve the temporal integrity of vessel movements, complete voyages identified by the ShipVoyage variable were kept within the same subset, preventing observations from a single voyage from being distributed across both training and testing datasets. This strategy minimised information leakage and provided a more realistic evaluation of model performance.

Within the 80% training partition, a further 80/20 split (scikit-learn's GroupShuffleSplit, also grouped by ShipVoyage) separated the data used to fit each model from a held-out validation set used to compute the headline MAE, RMSE, and R^2 figures reported in Section 6. The 20% test partition set aside at the outset was not used for any tuning or preprocessing decision and was evaluated only once, using the final fitted model, to obtain the independent test-set metrics reported alongside the operational baselines in Section 6.

Model performance was further assessed using five-fold grouped cross-validation (scikit-learn's GroupKFold, grouped by ShipVoyage and confined entirely to the 80% training partition, never touching the held-out test set) to evaluate each algorithm's generalization capability across different data partitions. This complementary evaluation reduced the dependence on a single train-validation split and provided a more robust comparison of the predictive performance of KNN, Random Forest Regression, and Multilayer Perceptron.

All experiments were implemented in Python 3.11.1 using the Anaconda environment and Jupyter Notebook 7.2.2. Data manipulation was performed using *pandas 2.2.3* and *NumPy 1.26.4 *, while *Scikit-learn 1.2.1* was used for model implementation and evaluation. Geospatial processing and visualisation were supported by *GeoPandas 1.0.1*, *Cartopy 0.23.0*, *Matplotlib 3.9.2*, and *Folium 0.17.0*. To improve computational efficiency when processing the AIS dataset, the data were read sequentially in batches of 100,000 records using the *chunksize* parameter, reducing memory requirements while preserving the integrity of the training workflow.

5.4. Performance Evaluation

The predictive performance of the three machine learning models was evaluated using complementary metrics that assess accuracy, goodness of fit, generalisation capability and computational efficiency. Since ETA forecasting is formulated as a regression problem, the evaluation focused on metrics commonly adopted for continuous prediction tasks [5,8].

Model accuracy was assessed using the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). MAE measures the average magnitude of prediction errors regardless of their direction, providing an intuitive estimate of the expected forecasting error. RMSE assigns greater weight to larger prediction errors, making it particularly useful for identifying models that occasionally produce substantial deviations from the observed arrival times.

The coefficient of determination (R^2) was used to evaluate the proportion of variance in the observed Remaining Travel Time (RTT) explained by each model. In addition to the validation R^2, the mean cross-validation R^2 was calculated to assess each algorithm's ability to generalise across different data partitions, thereby reducing the influence of a single train-test split and providing a more robust estimate of predictive performance.

Finally, execution time was recorded for each algorithm to assess computational efficiency. Although forecasting accuracy represents the primary evaluation criterion, computational performance is also relevant in operational port environments, where ETA predictions may need to be updated frequently using continuously received AIS data.

Together, these complementary metrics provide a comprehensive framework for comparing the predictive capability, robustness, and practical applicability of the selected machine learning models.

6. Results

6.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to examine the statistical properties of the final modelling dataset and to identify potential relationships among the selected predictor variables before evaluating the forecasting models. The analysis focused on feature correlations, multicollinearity, and variable distributions, providing additional insights into the engineered dataset's characteristics.

The correlation matrix presented in Figure 10 shows several meaningful relationships between the predictor variables and the target variable. As expected, the strongest positive correlation was observed between Remaining Travel Distance (RTD) and Remaining Travel Time (RTT) (0.69), confirming that vessels located farther from the destination generally require more time to complete their voyages. A weak positive correlation was also identified between Speed over Ground (SOG) and Course over Ground (COG) (0.12), suggesting a modest tendency for vessels travelling at higher speeds to maintain more consistent navigation courses.

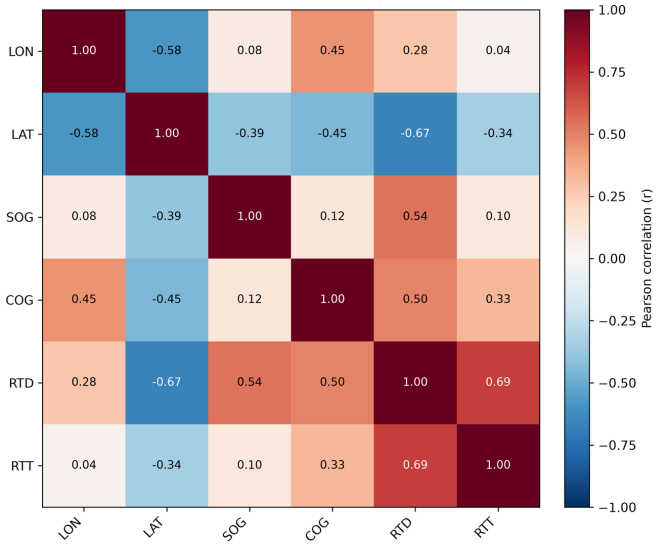


Figure 10. Correlation matrix (unscaled features).

Weak or negative correlations involving geographical coordinates were not considered problematic because latitude and longitude represent complementary spatial dimensions rather than independent explanatory variables. Similarly, variables such as MMSI and IMO exhibited negligible correlations with RTT, indicating that vessel identifiers contribute little to ETA prediction and therefore were not retained as predictor variables.

To further evaluate redundancy among the selected variables, multicollinearity was assessed using the Variance Inflation Factor (VIF). As illustrated in Figure 11, all predictor variables presented low-to-moderate VIF values, well below the conventional threshold of 5: LON (1.77), LAT (2.67), SOG (1.49), COG (1.59), and RTD (2.64). Although latitude and longitude jointly describe vessel position, their VIF values do not indicate problematic collinearity in this corrected dataset, and both were retained because they are essential for calculating Remaining Travel Distance (RTD). Overall, these results indicate that multicollinearity is not a concern for the selected set of predictors.

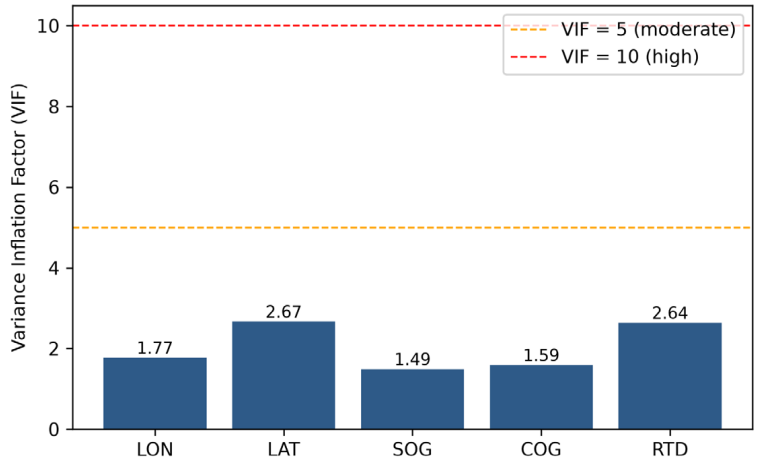


Figure 11. Variance Inflation Factor (VIF) for the selected predictor variables.

The distribution of the principal predictor variables is shown in Figure 12. Speed over Ground (SOG) values were concentrated below 20 knots (mean 9.6, median 10.3 knots; skewness −0.15), reflecting the operational characteristics of container vessels approaching the Port of Sines; the distribution is notably bimodal, with a pronounced cluster near 0–2 knots corresponding to manoeuvring or stationary vessels close to port, separate from a second cluster around 10–12 knots corresponding to normal cruising approach. Course over Ground (COG) showed multiple peaks (mean 148.2, median 131.5 degrees; skewness 0.29), indicating preferred navigation routes. Both RTD (mean 83.4, median 45.0 km; skewness 1.16) and RTT (mean 8.0, median 6.3 h; skewness 1.17) exhibited positively skewed distributions, consistent with the progressive reduction in both remaining distance and travel time as vessels approached the port; RTD additionally showed a secondary peak around 130–150 km, plausibly associated with vessels following specific approach routes from the Mediterranean (Section 4.1).

Feature scaling and logarithmic transformations were also examined based on the observed distributions. As detailed in Section 6.2, scaling did not have a uniform effect across models: it left Random Forest Regression performance essentially unchanged, had a small negative effect on KNN, and increased the variance of Multilayer Perceptron results across cross-validation folds. The original (unscaled) feature values were retained as the primary reported condition. The influence of feature scaling on forecasting performance is examined in detail in the following section.

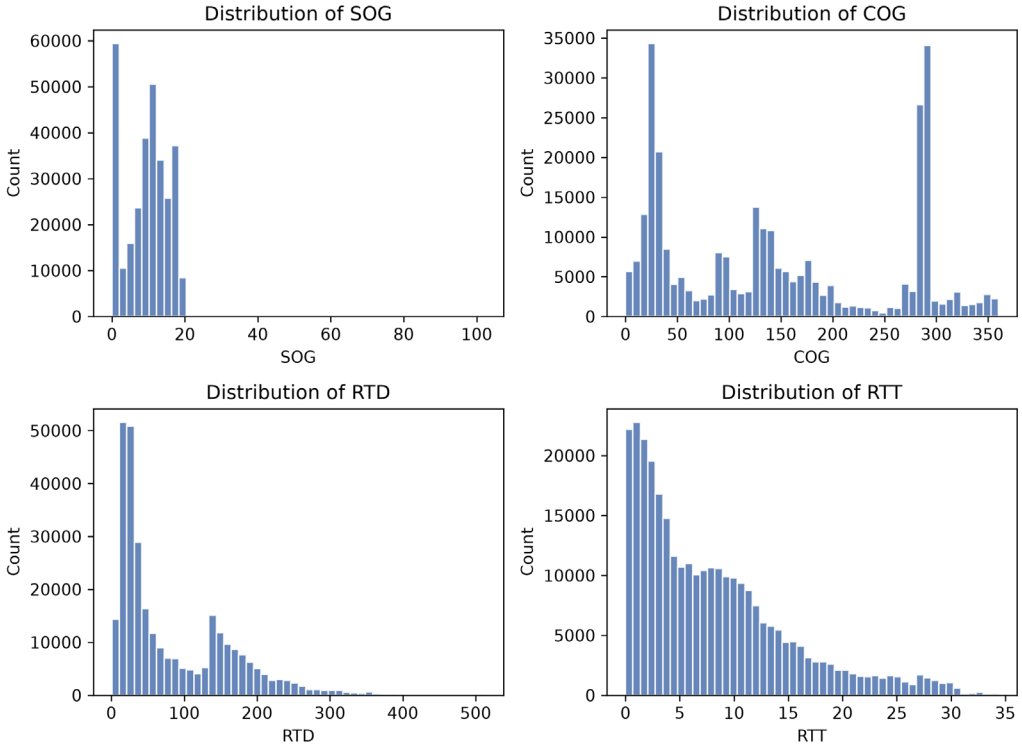


Figure 12. Distribution of principal predictor/target variables.

6.2. Effect of Feature Scaling

Feature scaling is commonly applied in machine learning to reduce differences in variable magnitude and improve model convergence. However, its effectiveness depends on the characteristics of the dataset and the learning algorithms employed. To investigate whether feature scaling improved ETA forecasting performance, the selected predictor variables were evaluated using both scaled and unscaled datasets.

All three models were retrained on an identical, leakage-free protocol (grouped train/validation/test partitions and grouped cross-validation, Section 5.3), changing only whether a StandardScaler (fit on the training data only, per split and per fold) was applied.

As shown in Figure 13, the effect of scaling was not uniform across models. Random Forest Regression results were essentially unchanged (mean cross-validation R^2 of 0.571 unscaled versus 0.570 scaled), K-Nearest Neighbours showed a small negative effect (0.449 versus 0.434), and the Multilayer Perceptron showed a sharp drop in a single cross-validation fold (R^2 of −8.24, against 0.10–0.87 in the remaining four folds), pulling its mean cross-validation R^2 from 0.305 (unscaled) to −1.478 (scaled).

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
KNN	0.670	0.818	0.095	0.389	0.201	0.434	0.256
RFR	0.751	0.870	0.381	0.388	0.460	0.570	0.202
MLP	0.619	0.607	-0.543	0.163	-8.236	-1.478	3.405

Figure 13. Fold-wise cross-validation R^2 obtained using scaled predictor variables. The cell highlighted in red (MLP, Fold 5) marks the anomalously low R^2 discussed in the text, attributable to scaling-induced instability in MLP training on a small per-fold sample.

To investigate the cause of the observed difference in the Multilayer Perceptron, the correlation structure of the scaled dataset was examined. Figure 14 shows that feature scaling did not measurably alter the correlation structure among predictor variables: the

scaled correlation matrix is numerically identical to the unscaled matrix shown in Figure 10 (maximum absolute difference below 10^{-6}), as expected, because Pearson correlation is mathematically invariant to the linear rescaling applied by a standard scaler.

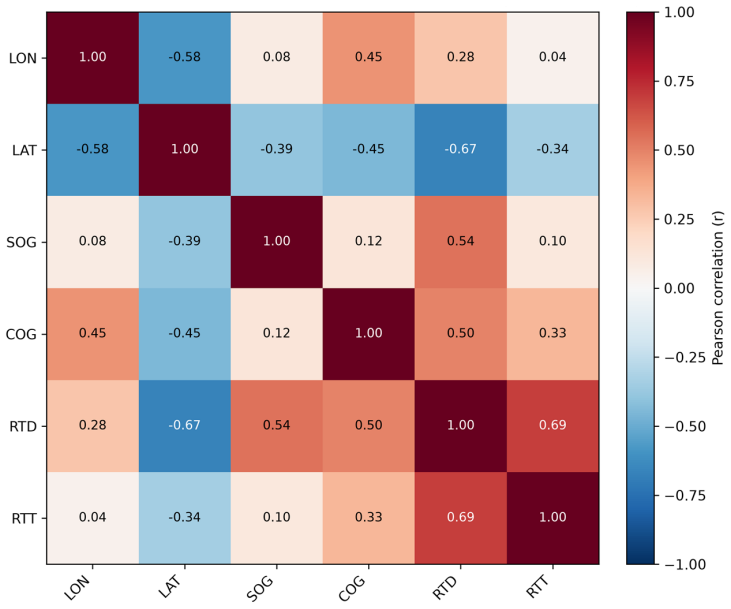


Figure 14. Correlation matrix (scaled features/StandardScaler).

This rules out a change in variable relationships as the explanation for the Multilayer Perceptron's degraded scaled performance. Random Forest Regression's near-identical scaled and unscaled results are consistent with this invariance, since tree-based ensembles partition the feature space using per-variable thresholds that are unaffected by monotonic linear rescaling.

As scaling provided no measurable benefit for any of the three models and introduced additional fold-to-fold variance for the Multilayer Perceptron, the original (unscaled) feature values were adopted as the primary reported condition for all subsequent analyses. Figure 15 presents the corresponding performance metrics for the three models under this unscaled condition, which are used throughout the remainder of this paper.

Model	Scaling	Mean CV R^2	Test MAE (h)	Test RMSE (h)	Test R^2
KNN	No	0.449	5.21	7.27	0.173
KNN	Yes	0.434	5.29	7.38	0.148
RFR	No	0.571	5.12	7.12	0.208
RFR	Yes	0.570	5.11	7.10	0.212
MLP	No	0.305	5.58	7.46	0.130
MLP	Yes	-1.478	5.66	7.92	0.019

Figure 15. Scaled vs. unscaled performance comparison for all three models. The row highlighted in red (MLP, Scaling = Yes) corresponds to the anomalous cross-validation fold discussed in the text (Section 6.2), reflecting scaling-induced instability in MLP training on a small per-fold sample.

These findings indicate that the effect of feature scaling on ETA forecasting using AIS data depends on the learning algorithm rather than being universally beneficial or harmful. Scaling is neutral for Random Forest Regression, mildly detrimental for K-Nearest Neighbors on this feature set, and a source of additional training variance-rather than a

systematic improvement or degradation-for the Multilayer Perceptron, consistent with the known sensitivity of gradient-based training to weight initialisation on small per-fold samples. Consequently, all subsequent analyses and model comparisons were performed using the unscaled dataset, which is adopted as the primary reported condition for the remainder of this paper.

6.3. Model Performance Comparison

The predictive performance of the three machine learning models was evaluated using the unscaled dataset, which demonstrated superior generalisation capability during the preliminary experiments. Performance was assessed using the mean cross-validation coefficient of determination (Mean CV R^2), validation R^2, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and execution time.

Overall, the three algorithms produced accurate ETA predictions, although noticeable differences were observed in their predictive capability, computational efficiency and generalisation performance. Table 5 summarises the execution times for each model, while Table 6 presents the corresponding performance metrics.

Table 5. Execution time of the evaluated machine learning models.

Model	Execution Time
KNN	0.22 s
RFR	55.95 s
MLP	7 min 24 s

Table 6. Performance comparison of the evaluated machine learning models.

Model	Mean CV R^2	Validation R^2	MAE	RMSE
KNN	0.449	0.510	3.061	4.692
RFR	0.571	0.639	2.692	4.026
MLP	0.305	0.504	3.438	4.720

As shown in Table 5, KNN required the shortest execution time, completing model training and prediction in a fraction of a second. Random Forest Regression also trained quickly, while the Multilayer Perceptron was, by a substantial margin, the most computationally demanding algorithm, consistent with the iterative, gradient-based nature of its training procedure.

The results indicate that Random Forest Regression achieved the highest cross-validation performance (mean CV R^2 = 0.571), demonstrating the most consistent generalisation across different data partitions. The Multilayer Perceptron produced the lowest cross-validation performance (0.305) and, when features were scaled, exhibited a sharp drop in a single fold, discussed in Section 6.2. K-Nearest Neighbors occupied an intermediate position (0.449) between the two.

Regarding prediction errors on the held-out validation set, Random Forest Regression achieved the lowest MAE and RMSE, while K-Nearest Neighbors and the Multilayer Perceptron produced somewhat higher, broadly comparable error levels. Section Held-Out Test-Set Performance, Operational Baselines and Robustness Checks additionally reports performance on a fully held-out test set of voyages, never used for tuning or cross-validation, to assess how these validation-stage results generalise to unseen data.

Held-Out Test-Set Performance, Operational Baselines and Robustness Checks

The results reported above (Table 6) correspond to the held-out validation set defined in Section 5.3. To assess how these results generalise to genuinely unseen data, Table 7

reports performance on the independent 20% test partition of voyages, which was set aside at the outset and evaluated only once, using each model's final fitted parameters.

Table 7. Performance of the evaluated machine learning models on the held-out test-set.

Model	Test MAE (h)	Test RMSE (h)	Test R^2	Macro MAE (h)	Macro RMSE (h)
KNN	5.21	7.27	0.173	3.24	4.06
RFR	5.12	7.12	0.208	3.08	3.79
MLP	5.58	7.46	0.130	3.32	4.09

Random Forest Regression achieved the lowest test-set MAE and RMSE and the highest test-set R^2, consistent with its performance on the validation set. All three models showed somewhat lower R^2 on the test set than on the validation set, which is expected given the small number of voyages available for testing (n = 15) and is discussed further as a limitation in Section 8.

Additionally, distinguished between two ways of aggregating record-level errors into a single performance figure. The results presented above follow the conventional, record-pooled ("micro-averaged") approach, in which all AIS records from all test voyages are combined before computing MAE and RMSE; this implicitly gives more weight to longer or more densely sampled voyages. Table 8 additionally reports "macro-averaged" metrics, computed by first calculating MAE and RMSE separately for each of the 15 test voyages and then averaging across voyages with equal weight, consistent with the view that the voyage, not the individual AIS message, is the effectively independent unit of observation in this dataset (Section 8).

Table 8. Micro- vs. macro-averaged test-set error metrics.

Model	Micro MAE (h)	Macro MAE (h)	Micro RMSE (h)	Macro RMSE (h)	SD Across Voyages (MAE)
KNN	5.29	3.24	7.38	4.06	3.08
RFR	5.12	3.08	7.12	3.79	2.96
MLP	5.66	3.32	7.92	4.09	3.31

Macro-averaged errors are consistently lower than the micro-averaged values reported above for all three models. This is explained by a single long, densely sampled test voyage (25,633 records, approximately 30% of all test records) with a comparatively poor fit, which disproportionately dominated the pooled (micro) average; under macro-averaging, it counts as one voyage among fifteen rather than 25,633 individual records.

The standard deviation of per-voyage MAE is comparable in magnitude to the mean itself for all three models, confirming that a single mean value conceals substantial voyage-to-voyage variability. The same two voyages rank as the most difficult to predict across all three models regardless of algorithm, suggesting that this difficulty is a property of those specific voyages (e.g., atypical navigational behaviour) rather than a weakness specific to any one model. Importantly, the relative ranking of the three models is unchanged under macro-averaging: Random Forest Regression remains the best-performing model by both MAE and RMSE. The macro-averaged (voyage-level) metrics were adopted as the primary performance measure in this revision, while retaining the micro-averaged values for transparency and comparability with the previous version.

To characterise the uncertainty of the cross-validation estimates reported in Table 9, the individual fold-wise R^2 scores along with a 95% confidence interval for the mean, computed using a t-distribution with 4 degrees of freedom given the small number of folds (n = 5) were also reported.

Table 9. Fold-wise cross-validation R^2 and 95% confidence intervals.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean CV R^2	95% CI
KNN	0.715	0.790	0.170	0.370	0.203	0.450	[0.092, 0.807]
RFR	0.751	0.870	0.387	0.388	0.458	0.571	[0.292, 0.850]
MLP	0.775	−0.104	0.381	0.275	0.197	0.305	[−0.091, 0.701]

The width of these intervals—and the fact that the interval for the Multilayer Perceptron includes negative values—makes explicit the substantial uncertainty inherent in estimating generalisation from only five voyage-grouped folds and reinforces our recommendation (Section 6.4) that Random Forest Regression, rather than the Multilayer Perceptron, be treated as the primary model given the current dataset size.

Finally, to assess whether forecasting accuracy depends on a vessel's distance from the port, Table 10 reports test-set MAE (Random Forest Regression) stratified by Remaining Travel Distance (RTD).

Table 10. Test-set MAE by Remaining Travel Distance (RTD) bin, Random Forest Regression.

RTD Bin	*n* Records	MAE (h)
0–10 km	2068	0.12
10–30 km	29,666	4.62
30–60 km	19,914	4.46
60–100 km	8079	5.86
100–200 km	20,659	6.09
200+ km	5424	7.40

Error is lowest immediately before arrival, as expected since the remaining time itself is short and bounded close to port, and generally increases with remaining distance, confirming that longer-horizon predictions are harder; the same qualitative pattern is observed for KNN and MLP.

To assess the practical value of the proposed machine learning workflow relative to simple operational alternatives, three baselines were evaluated on the same held-out test voyages: a kinematic baseline (remaining travel time estimated as RTD divided by current SOG), a historical-average (climatology) baseline (mean RTT observed in the training voyages among records at a similar RTD, using 10 km bins), and the AIS-reported ETA field itself, converted into a remaining-time estimate and compared against the ATA-based RTT. Table 11 presents the results.

Table 11. Comparison of the machine learning workflow against operational baselines on the held-out test set.

Method	Test MAE (h)	Test RMSE (h)	Test R^2
Kinematic baseline (RTD/SOG)	10.69	25.41	−9.09
AIS-reported ETA	7.00	9.70	−0.30
Historical average (10 km bins)	5.57	7.21	0.19
Random Forest Regression (best model)	5.12	7.12	0.21

All three non-learning baselines perform markedly worse than the machine learning models. Notably, the AIS-reported ETA field—arguably the current operational status quo at the port—also produced a negative R^2 (worse than always predicting the mean RTT), computed over the subset of test records for which a non-missing, non-expired reported ETA was available. The historical-average baseline is the strongest of the three non-learning

baselines, performing close to, though still below, the machine learning models; this provides a more conservative and more informative picture of the added value of the proposed workflow than a comparison against the kinematic baseline alone.

Two additional robustness checks were performed on Random Forest Regression, the best-performing model. First, a chronological split (training on the 60 earliest voyages, testing on the 15 most recent) gave a lower test-set error than the random grouped split (MAE = 2.04 h, RMSE = 3.28 h, R^2 = 0.518, versus 5.12 h, 7.12 h and 0.208 respectively); however, the kinematic baseline improved by a similar proportion on the same chronological test voyages (MAE = 4.33 h, versus 10.69 h under the random split), suggesting that this improvement most likely reflects the composition of the most recent voyages rather than a genuine temporal-generalisation effect. Given the small number of test voyages (n = 15), this analysis is reported as indicative and exploratory rather than as a robust estimate of temporal drift. It was performed only for Random Forest Regression, given the computational cost of repeating full grouped cross-validation for all three models within the available time.

Second, a sensitivity analysis was performed by re-including the five loitering voyages excluded in Section 4.3 (80 voyages in total). Cross-validation and test-set R^2 deteriorated sharply (mean CV R^2 = −0.199; test R^2 = −0.099), while test-set MAE increased only marginally (5.15 h versus 5.12 h); the 90th and 95th percentile absolute errors increased substantially (9.22 h and 13.27 h, versus 6.87 h and 7.89 h without loitering voyages). This pattern indicates that the model continues to perform similarly well on most typical test observations when loitering voyages are included, but produces a small number of much larger errors on the atypical, loitering-affected trajectories, to which RMSE and R^2 are considerably more sensitive than MAE. This provides direct empirical support for treating loitering behaviour as a distinct operational regime, discussed further in Sections 7 and 8, rather than folding it into the main forecasting workflow.

To facilitate the practical interpretation of these results, Table 12 summarises the principal strengths, limitations, and recommended application scenarios for each model.

Table 12. Comparative assessment of the evaluated machine learning models.

Model	Main Strengths	Main Limitations	Recommended Application
KNN	Fastest execution by a wide margin; simple implementation; high interpretability	Lowest generalisation capability of the three models; tail errors increase further under feature scaling	Real-time applications and rapid deployment where computational cost is the primary constraint
RFR	Highest cross-validation and test-set performance; lowest MAE/RMSE; invariant to feature scaling	Lower interpretability than linear models; requires more compute than KNN, though still fast in absolute terms	High-accuracy ETA forecasting and operational decision support; recommended primary model under current dataset size
MLP	Capable of modelling complex nonlinear interactions among predictor variables	Lowest cross-validation performance observed; substantially longer training time; sensitive to feature scaling and weight initialisation on small per-fold samples	Secondary role pending further tuning or additional training data; not recommended as the primary model at the current dataset size

Overall, the comparison demonstrates that each algorithm offers distinct advantages depending on the operational requirements. Random Forest Regression achieved the highest predictive performance and the most consistent generalisation, without sensitivity to feature scaling; KNN remains an attractive solution for applications where computational efficiency is the primary concern, at some cost in generalisation capability; and MLP, while capable of modelling complex nonlinear interactions, showed the lowest generalisation performance and the longest training times under the present dataset size.

6.4. Practical Implications for Model Selection

The comparative evaluation demonstrates that no single machine learning model is optimal for every operational context. Instead, the most appropriate algorithm depends on the balance between forecasting accuracy, computational efficiency, and the specific requirements of port operations.

Random Forest Regression (RFR) achieved the highest overall predictive performance and demonstrated the most consistent generalisation capability, making it the most suitable model for applications where forecasting accuracy is the primary objective. Such applications include berth planning, resource allocation, and operational decision support, where small improvements in ETA prediction can contribute to more efficient coordination among port stakeholders. Because tree-based ensembles are invariant to linear feature scaling, RFR also offers this performance without requiring a scaling decision, simplifying deployment.

The Multilayer Perceptron (MLP) was capable of modelling complex nonlinear interactions among predictor variables but produced the lowest cross-validation and test-set performance of the three models in this study, together with substantially longer training times and greater sensitivity to feature scaling and weight initialisation. Under the current dataset size, MLP is therefore not recommended as the primary model; it may become more competitive with additional training data or further hyperparameter tuning, which was identified as a direction for future work.

Although K-Nearest Neighbors (KNN) achieved lower cross-validation performance than Random Forest Regression, it required substantially less computational time than either alternative and produced competitive prediction errors. Consequently, KNN may represent a suitable solution for applications requiring rapid model execution or operating under limited computational resources, at some cost in generalisation capability relative to RFR.

These findings indicate that model selection should consider not only predictive performance but also operational requirements and implementation constraints. Consequently, the choice of an ETA forecasting model should be aligned with the intended decision-support application rather than based exclusively on a single performance metric.

7. Discussion

7.1. Comparison with Previous ETA Forecasting Studies

The findings of this study are broadly consistent with previous research demonstrating the potential of machine learning for improving ETA forecasting based on AIS data. Similar to earlier studies, the results confirm that machine learning algorithms can accurately estimate vessel arrival times by exploiting navigational information derived from AIS transmissions, supporting the increasing adoption of data-driven approaches in maritime logistics [5,8].

The comparative evaluation of KNN, Random Forest Regression, and Multilayer Perceptron also agrees with the trends reported in the literature. Valero et al. [8] identified Random Forest as one of the most effective algorithms for ETA prediction, highlighting its robustness when modelling nonlinear relationships between navigational variables. Likewise, Kolley et al. [5] and Magalhães [6] demonstrated that KNN and MLP constitute reliable alternatives for predicting Remaining Travel Time using AIS-derived features. The present study confirms that all three approaches provide high predictive capability when applied to short-sea ETA forecasting, while also revealing differences in their generalisation performance, computational efficiency and operational suitability.

An important contribution of this study concerns the operational context in which the models were evaluated. Many published ETA forecasting studies focus on long-distance voyages, where meteorological conditions, ocean currents, and route optimisation signifi-

cantly influence prediction accuracy [8,9]. In contrast, the present research demonstrates that accurate ETA prediction can be achieved for short-sea shipping using exclusively AIS-derived variables, without incorporating additional environmental data. This finding suggests that, for short-range approaches to port terminals, vessel trajectory information may provide sufficient predictive information to support operational decision-making.

Furthermore, this study extends prior research by evaluating the combined influence of AIS data preprocessing and machine learning model selection within a single experimental framework. While earlier studies have primarily focused on comparing forecasting algorithms, the results presented here indicate that the quality of the modelling dataset and the engineering of AIS-derived features are equally important determinants of forecasting performance. This broader perspective contributes to the development of reproducible ETA forecasting workflows that integrate data preparation and model selection as complementary components of the prediction process.

7.2. Implications for ETA Forecasting

The results demonstrate that ETA forecasting performance depends not only on the choice of the machine learning algorithm but also on the quality of the underlying dataset and the decisions made during data preparation. This finding reinforces the view that predictive performance is the outcome of an integrated modelling pipeline in which data preprocessing, feature engineering, and model selection jointly contribute to forecasting accuracy.

An important finding concerns the influence of feature scaling, examined in detail in Section 6.2. Feature scaling did not have a uniform effect across the three algorithms: it left Random Forest Regression performance essentially unchanged, since tree-based ensembles partition the feature space using thresholds that are unaffected by a monotonic linear rescaling; it had a small negative effect on K-Nearest Neighbours; and it increased cross-validation variance for the Multilayer Perceptron, consistent with the known sensitivity of gradient-based training to weight initialisation on small per-fold samples, rather than with any change in the underlying relationships among variables (the scaled correlation matrix, Figure 14, is numerically identical to the unscaled matrix in Figure 10, confirming that Pearson correlation is invariant to this linear rescaling). The original, unscaled magnitudes of variables such as Speed over Ground (SOG), Course over Ground (COG), and Remaining Travel Distance (RTD) were nonetheless adopted as the primary reported condition throughout this paper, since they produced the most consistent results across all three models. This observation is consistent with the findings of Kolley et al. [5] and Magalhães [6], who also reported strong predictive performance when AIS-derived variables were maintained in their original scale, although our results indicate that this benefit is algorithm-dependent rather than universal.

The study also highlights the importance of feature engineering for ETA forecasting. Rather than relying exclusively on raw AIS observations, the inclusion of derived variables, such as Remaining Travel Distance, enabled the models to incorporate information about the progression of each voyage. This finding supports previous research indicating that carefully engineered navigational features often contribute more to forecasting performance than increasing model complexity alone [5,8].

Another important implication concerns the interpretation of model performance. Although the evaluated algorithms achieved comparable validation accuracy, differences emerged in their ability to generalise across different data partitions and in their computational requirements. This indicates that selecting an ETA forecasting model should not rely exclusively on a single performance metric. Instead, forecasting systems intended for

operational deployment should balance predictive accuracy, robustness, and computational efficiency according to the application's requirements.

Overall, these findings suggest that improvements in ETA forecasting are achieved through the combined optimisation of data preparation and model development rather than through algorithm selection alone. For short-sea shipping applications, a carefully designed preprocessing workflow, combined with appropriate feature engineering, may contribute as much to forecasting performance as the choice of predictive model.

7.3. Operational Implications for Ports

Accurate ETA forecasting has implications that extend beyond predictive performance, directly supporting operational planning and decision-making within port environments. Reliable arrival time estimates enable ports to anticipate vessel movements more effectively, improving the coordination of nautical services, berth allocation, terminal operations, and resource scheduling. In short-sea shipping, where vessels often arrive within relatively narrow time windows, even small improvements in ETA accuracy can contribute to more efficient operational planning and reduced waiting times.

The findings of this study indicate that machine learning models trained exclusively with AIS-derived information can provide accurate ETA predictions without requiring additional external data sources. This characteristic simplifies implementation in operational environments because AIS data are continuously available and already integrated into the digital infrastructure of most commercial ports. Consequently, the proposed forecasting workflow can be incorporated into existing port information systems to support real-time operational decision-making.

The comparative evaluation of the three machine learning models also demonstrates that different operational contexts may require different forecasting solutions. Random Forest Regression is particularly suitable for applications where maximising forecasting accuracy is the primary objective, such as berth planning and terminal coordination, and offers this performance without the added complexity of a feature-scaling decision. K-Nearest Neighbors, owing to its very low computational requirements, represents an attractive option for applications requiring rapid predictions or operating under limited computational resources, at the cost of generalisation capability. The Multilayer Perceptron, while capable of modelling complex nonlinear interactions, showed the lowest generalisation performance and the longest training times in this study, and its adoption for operational deployment would likely require additional training data or further tuning.

Beyond the selection of individual algorithms, the results emphasise that the reliability of ETA forecasting systems depends on the quality of the complete modelling workflow. The structured preprocessing of AIS data, the construction of operationally meaningful features, and the selection of appropriate predictive models together provide a robust foundation for supporting port operations. This integrated perspective aligns with the ongoing digital transformation of ports, where data-driven forecasting is increasingly recognised as a key capability to improve operational efficiency, resource utilisation and service reliability.

Overall, the proposed workflow demonstrates that reproducible machine learning approaches based on routinely available AIS data can support more informed operational decisions while requiring only limited additional data collection. As ports continue to adopt digital decision-support systems, methodologies that integrate robust data preparation with predictive modelling are likely to play an increasingly important role in enhancing maritime logistics performance.

8. Conclusions

This study investigated how AIS data preprocessing and machine learning model selection jointly influence ETA forecasting performance for short-sea shipping, using the Port of Sines as an empirical case study. A reproducible forecasting workflow was developed that integrates dataset construction, feature engineering, machine learning model development, and performance evaluation into a single methodological framework.

The results demonstrate that accurate ETA forecasting depends not only on the selection of an appropriate machine learning algorithm but also on the quality of the underlying modelling dataset. Structured data preprocessing, voyage identification, and feature engineering substantially enhanced the predictive capability of the evaluated models, underscoring the importance of treating dataset construction as an integral part of the forecasting process rather than merely a preliminary technical stage. In particular, the results showed that the effect of feature scaling was algorithm-dependent: essentially neutral for Random Forest Regression, mildly negative for K-Nearest Neighbors, and a source of additional training variance for the Multilayer Perceptron. The original, unscaled magnitude of AIS-derived navigational variables was adopted as the primary reported condition throughout this study, as it produced the most consistent results across all three models.

The comparative evaluation of K-Nearest Neighbors, Random Forest Regression, and Multilayer Perceptron confirmed that each algorithm presents distinct advantages depending on the intended operational application. Random Forest Regression achieved the strongest overall generalisation performance and the lowest prediction errors, on both the held-out validation set and an independent held-out test set, and did so without sensitivity to feature scaling; K-Nearest Neighbours offered substantially lower computational requirements at some cost in generalisation capability; and the Multilayer Perceptron, while capable of modelling complex nonlinear interactions, showed the lowest generalisation performance and the longest training times under the present dataset size. These findings indicate that machine learning model selection should consider operational requirements alongside conventional accuracy metrics.

From a methodological perspective, this study contributes to the ETA forecasting literature by demonstrating that improvements in predictive performance arise from the combined optimisation of data preprocessing, feature engineering and model selection rather than from algorithm selection alone. The proposed workflow provides a transparent and reproducible framework that can support future research using AIS data for maritime forecasting applications.

From a practical perspective, the proposed methodology can support digital decision-making in ports by improving the reliability of ETA predictions used for berth planning, resource allocation, and operational coordination. Because the workflow relies exclusively on routinely available AIS data, it can be readily integrated into existing port information systems without requiring additional external data sources, making it particularly suitable for short-sea shipping operations.

This study is subject to some limitations. The experimental evaluation was conducted using AIS data from a single port and focused on container vessel traffic approaching the Port of Sines, over a final modelling dataset of 75 voyages; given this small effective sample size, some of the robustness checks reported in Section Held-Out Test-Set Performance, Operational Baselines and Robustness Checks (the chronological split and the loitering-sensitivity analysis) are necessarily indicative and exploratory rather than statistically conclusive, and the chronological check was performed for the best-performing model (Random Forest Regression) only, owing to the computational cost of repeating full grouped cross-validation for all three models within the available time. Vessel-level generalisation

(i.e., performance on vessels not seen during training) was not separately tested, since some vessels contribute more than one voyage to the dataset.

Furthermore, only three supervised machine learning algorithms were evaluated, and external variables such as weather conditions, sea state, and tidal information were intentionally excluded to isolate the influence of AIS-derived features and preprocessing decisions. Finally, voyages exhibiting pronounced loitering behaviour were excluded from the main modelling dataset; Section Held-Out Test-Set Performance, Operational Baselines and Robustness Checks shows that including them substantially increases tail error (P90/P95) and reduces R^2 while leaving MAE largely unchanged, indicating that the proposed workflow, as currently trained, is not well suited to this atypical operational regime.

Future research should evaluate the proposed workflow across different ports, vessel categories, and geographical contexts, using larger AIS datasets that would support vessel-level (leave-one-vessel-out) and more statistically robust temporal validation than was possible with the present sample, while investigating the contribution of additional data sources and more advanced forecasting techniques. A dedicated anomaly-aware sub-model for loitering and other irregular manoeuvring behaviour, rather than exclusion from the main workflow, is also identified as a promising direction. Comparative studies involving hybrid machine learning models, deep learning approaches, and real-time operational implementation may further enhance ETA forecasting performance and contribute to the development of intelligent decision-support systems for maritime logistics.

Generative artificial intelligence (GenAI) tools were used during the preparation of this manuscript to assist with drafting, restructuring, and linguistic refinement of the text, as well as generating selected conceptual figures illustrating the methodological workflow. All scientific content, including the study design, data collection, data preprocessing, model development, analyses, interpretation of results and conclusions, was conceived, conducted, verified, and critically reviewed by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the content presented in this manuscript.

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

Funding: Content produced within the scope of the Agenda "NEXUS-Pacto de Inovação–Transição Verde e Digital para Transportes, Logística e Mobilidade", financed by the Portuguese Recovery and Resilience Plan (PRR), with no. C645112083–00000059 (investment project no. 53).

Data Availability Statement: The code can be accessed through https://github.com/marreirosj/RTT-prediction-with-ML, accessed on 29 July 2026.

Acknowledgments: The authors wish to thank: APS-Administração dos Portos de Sines e do Algarve, for providing access to the data and resources that made this study possible. NECE-Research Centre for Business Sciences funded by the Multiannual Funding Program of R&D Centers of FCT-Fundação para a Ciência e Tecnologia, Portugal, under Grant number UID/04630/2025, DOI: 10.54499/UID/04630/2025. CIEQV–Life Quality Research Center, financed by national funds through FCT–Foundation for Science and Technology, I.P., under the project n. UID/CED/04748/2025.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Côté, P.O.; Nikanjam, A.; Ahmed, N.; Humeniuk, D.; Khomh, F. Data Cleaning and Machine Learning: A Systematic Literature Review. *Autom. Softw. Eng.* **2024**, *31*, 54. [CrossRef]
2. Balster, A.; Hansen, O.; Friedrich, H.; Ludwig, A. An ETA Prediction Model for Intermodal Transport Networks Based on Machine Learning. *Bus. Inf. Syst. Eng.* **2020**, *62*, 403–416. [CrossRef]
3. Park, K.; Sim, S.; Bae, H. Vessel estimated time of arrival prediction system based on a path-finding algorithm. *Marit. Transp. Res.* **2021**, *2*, 100012. [CrossRef]
4. Svanberg, M.; Santén, V.; Hörteborn, A.; Holm, H.; Finnsgård, C. AIS in maritime research. *Mar. Policy* **2019**, *106*, 103520. [CrossRef]
5. Kolley, L.; Rückert, N.; Kastner, M.; Jahn, C.; Fischer, K. Robust berth scheduling using machine learning for vessel arrival time prediction. *Flex. Serv. Manuf. J.* **2023**, *35*, 29–69.
6. Magalhães, A.M. Smart Berth Planning for Commercial Ports-Vessel Arrival Time Prediction and Berth Allocation Using Machine Learning. 2023. ISBN 6451120830000. Available online: https://estudogeral.uc.pt/bitstream/10316/107937/1/Afonso%20Matoso%20Magalh%C3%A3es.pdf (accessed on 29 July 2026).
7. Zhang, M.; Huang, L.; Wen, Y.; Zhang, J.; Huang, Y.; Zhu, M. Short-Term Trajectory Prediction of Maritime Vessel Using k-Nearest Neighbor Points. *J. Mar. Sci. Eng.* **2022**, *10*, 1939. [CrossRef]
8. Valero, C.I.; Martínez, Á.; Oltra, R.; Gil, H.; Boronat, F.; Palau, C.E. Prediction of the Estimated Time of Arrival of container ships on short-sea shipping: A pragmatical analysis. *IEEE Lat. Am. Trans.* **2022**, *11*, 2354–2362. [CrossRef]
9. Tsolaki, K.; Vafeiadis, T.; Nizamis, A.; Ioannidis, D.; Tzovaras, D. Utilizing machine learning on freight transportation and logistics applications: A review. *ICT Express* **2023**, *9*, 284–295. [CrossRef]
10. Gutierrez-Torre, A.; Berral, J.L.; Buchaca, D.; Guevara, M.; Soret, A.; David, C. Improving maritime traffic emission estimations on missing data with CRBMs. *Eng. Appl. Artif. Intell.* **2020**, *94*, 103793. [CrossRef]
11. Rice, J.R. The Algorithm Selection Problem. *Comput. Sci. Tech. Rep. Pap.* **1975**, *99*, 75–152.
12. Wolpert, D.H.; Macready, W.G. No Free Lunch Theorems for Optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 259–264. [CrossRef]
13. Hwang, H.S.; Hwang, T.M.; Youn, I.H. Efficacy evaluation of adaptive collision avoidance systems for autonomous maritime surface ships based on target ships' maneuvering behaviors. *J. Int. Marit. Saf. Environ. Aff. Shipp.* **2024**, *8*, 2298250. [CrossRef]
14. Burkov, A. *The Hundred-Page Machine Learning Book*; Andriy Burkov: Quebec City, QC, Canada, 2019; Volume 5, pp. 132–150.
15. Gholamy, A.; Kreinovich, V.; Kosheleva, O. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Int. J. Intell. Technol. Appl. Stat.* **2018**, *1209*, 105–111.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.