Enok Sanne Liland

Digital Twin For Resource Management In Norwegian Shipyards

Graduate thesis in Engineering and ICT
Supervisor: Jan Ola Strandhagen
Co-supervisor: Marco Semini
June 2023

Graduate thesis

NTNU
Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

NTNU
Norwegian University of
Science and Technology

Enok Sanne Liland

Digital Twin For Resource Management In Norwegian Shipyards

Graduate thesis in Engineering and ICT
Supervisor: Jan Ola Strandhagen
Co-supervisor: Marco Semini
June 2023

Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

NTNU
Norwegian University of
Science and Technology

Preface

This report is a Master's thesis in production management as part of the ICT & Operation Management study program at the Department of Mechanical and Industrial Engineering at the Norwegian University of Science and Technology. The study was conducted in the spring of 2023, and the initial motivation for the study was derived from personal interest and objectives given by a case company

I would like to thank my supervisors, Jan Ola Strandhagen and Marco Semini, professors at NTNU, for their great feedback and guidance during the semester. I would also like to thank the case company for their participation in the project and great communication.

Finally, I would like to extend my sincere gratitude to my fellow students for the great discussions and cheerful dialogues. A special thanks go to my family, girlfriend, and friends for their support and encouragement.

Summary

This master thesis explores the potential of digital twin technology to improve operational management in Norwegian shipyards. Amidst rising global competition and fluctuating demand in shipbuilding, Norwegian shipyards have been compelled to seek innovative solutions to enhance cost-efficiency and maintain their market position. This study focuses on digital twin technology as a promising approach for operational optimization, particularly in resource management. Two research questions are posed:

- RQ1: How can the concept of a digital twin be introduced and applied in Norwegian shipyards to improve their operations
- RQ2: How can digital twin technology influence decision-making processes in managing labor hours within shipyard operations?

To answer these research questions, the study employs an integrative approach combining a literature review, interviews with industry experts, and a case study focused on the estimation of labor hours for new potential projects using digital twin technology.

The research highlights the competitive pressures Norwegian shipyards face, particularly from regions with lower labor costs. It identifies the potential for Norwegian shipyards to be globally competitive in advanced segments if they can reduce costs by 10-15 percent. Implementing digital twin technology could be key to achieving this reduction, but it requires careful planning and phased implementation. This thesis suggests an incremental approach, where digital twins are built starting from a few applications, focused on high-impact operational areas, and then gradually evolving into a comprehensive model. In the current state of Norwegian shipyards, resource management emerges as the paramount domain for deployment, exhibiting substantial promise in terms of applicability and potential impact.

Specific domains identified for digital twin integration include resource management, predictive maintenance, disaster and emergency response, and crane and equipment utilization. Furthermore, the research found that digital twin technology could automate the labor hours estimation process, leading to more efficient and accurate predictions, better resource allocation, and enhanced cost management, which is crucial given the high labor costs in Norway.

However, the thesis acknowledges the challenges and prerequisites for successful implementation, such as data availability and quality, integration with existing systems, technical expertise, and investment willingness. The research recommends a modular, scalable, and continuously evaluated approach to implementing digital twin technology in shipyards.

In conclusion, this study provides structured insights and a framework for implementing digital twin technology in Norwegian shipyards. Through its incremental approach, it seeks to enhance operational efficiencies, decision-making capabilities, and cost management, given the prerequisites are adequately addressed. While it successfully addresses the research questions, it also acknowledges the evolving nature of digital twin technology and suggests vigilance and adaptability for future developments.

Table of Contents

List of Figures

List of Tables

1 Introduction

This chapter introduces the main topics of the master thesis. It starts by introducing the background and motivation of this master thesis. After that, the research questions are defined. Finally, it lays forward the project's scope and structure.

1.1 Background and motivation

The shipbuilding industry has played a significant role in the economy of Norway for many years (Amdam and Bjarnar 2015). It's been a long-established industry—reasons for it are due to geographical, political, and social factors. Shipyards are often central to local communities and provide employment and other economic benefits. Employing more than 80,000 individuals and boasting a turnover of 476 billion NOK in 2019, the maritime sector stands as the second-largest industry segment in Norway (Stensvold 2022). However, as global competition has increased, particularly from countries in East Asia and southern Europe, many Norwegian shipyards have faced challenges. The drop in oil prices from 2015 has contributed to low demand for vessels used in the oil and gas industry, which stood for about 80-90% of their production at its peak. To remain competitive, these shipyards have had to specialize in producing advanced and customized ships and vessels that meet the specific needs of their clients. This specialization has allowed them to remain viable in a highly competitive market (Semini, Brett, Hagen et al. 2018).

Norwegian shipyards concentrate mainly on outfitting and offshoring most of their steel work in Eastern European countries with lower labor costs and cheaper raw materials(Semini, Brett, Hagen et al. 2018). The general demand for ships has decreased globally. There is also a shift in demand from oil and gas–related projects to other markets, such as ferries, passenger and cruise ships, and ships for the offshore wind market (Strandhagen 2022). These markets tend to have lower profit margins than the oil and gas market, contributing to the decline in profitability for shipyards. This has contributed to more than 230 yards being closed down worldwide between 2014 and 2021 worldwide (Statista 2022).

The critical challenge, from an operations management point of view, involves cost-efficient operations. Operational management in shipyards is here used as a term for all administrative and physical tasks performed at the yard as a part of the production. Resyard (2023) state that Norwegian yards can still be globally competitive in the most advanced segments if they can reduce costs by 10-15 percent. Shipbuilding in Norway is an engineer-to-order operation and has only a few big projects in the year with low repetition (Hagen and Erikstad 2014). This project focuses on the potential applications of the digital twin concept in operational management in Norwegian shipyards.

Industry 4.0 is a technologically driven industrial revolution that is changing the manufacturing industry (Frank et al. 2019). The enabling industry 4.0 Technologies are expected to cause disruptive changes in manufacturing. Research on these technologies has accelerated the recent years. Studies mainly focus on universal assessments of Industry 4.0 of manufacturing, and research on the actual application of digital technologies in manufacturing still needs to be done (Zheng et al. 2021). These advanced technologies form the fundamental infrastructure of a digital twin, a virtual representation that mirrors the real-world object, system, or process. This infrastructure facilitates the simulation, analysis, and understanding of the performance and behavior of the corresponding physical entity, contributing significantly to predictive maintenance, operational efficiency, and system optimization (Kritzinger et al. 2018).

There is an emerging research stream on Industry 4.0 in engineer to order (ETO) (Cannas and Gosling 2021). The application of Industry 4.0 technologies in ETO manufacturing is a vital part of future research in the field (Zennaro et al. 2019). However, existing research has considered only a limited number of specific technological applications to particular areas or processes in yard operations (Strandhagen 2022). Despite this increasing interest, the exploration of digital twins technologies, specifically in shipyard operations, remains somewhat limited. This highlights a crucial gap in existing literature, underlining the need for more in-depth studies to comprehensively understand their potential role and benefits in shipyard settings.

Digital twin technology has emerged as a promising approach to enhance various aspects of product lifecycle management by creating a digital replica of a physical object, system, or process (Glaessgen and Stargel 2012). This digital representation allows for real-time monitoring, simulation, and analysis of the physical counterpart, facilitating improved decision-making, problem-solving, and performance optimization (Tao et al. 2018). The concept of digital twins has gained significant traction in recent years, particularly in the context of the Industry 4.0 paradigm, where the integration of cyber-physical systems, the Internet of Things (IoT), and advanced data analytics technologies play a crucial role in transforming industrial processes (Leng et al. 2020).

In the shipyard industry, digital models have been typically utilized from a product-oriented viewpoint, meaning they've been used mainly for virtual representation of the items being manufactured in the yard rather than for the yard itself. It's common to see CAD models of ships, windmills, platforms, and their components in sectors like shipbuilding and offshore construction. However, the methodologies for pursuing resource efficiency and flow in actual production are significantly more basic and incomplete, with a scarcity of digital twin applications, particularly in the yard industry (Resyard 2023). A theoretical foundation needs to be established for the creation, implementation, and exploitation of digital twins. These digital twins would aid in the visual and real-time planning of personnel, equipment, and infrastructure. Additionally, it's crucial to pinpoint the specific technologies that should be incorporated into a digital twin in this setting.

The benefits of digital twins in shipyard settings are manifold. Foremost among them is their potential to drastically improve operational efficiency. By creating a virtual replica of the shipyard, stakeholders can simulate different scenarios, foresee potential issues, and optimize resource allocation. This predictive capability could lead to significant reductions in operational costs, thereby increasing the competitiveness of the shipyards (Wang et al. 2022). The integration of IoT devices and sensors within the digital twin framework can further provide real-time data, enabling immediate response to emergent situations.

Nevertheless, the implementation of digital twins in shipyards is not without challenges. Given the substantial investments required for the technology's deployment, it's crucial to research to back it up and a good strategic plan. This Master's thesis will investigate how the concept of a digital twin can be implemented and utilized in Norwegian shipyards to enhance their operational performance.

Production costs are primarily influenced by investments and labor expenses. Given that materials and ship equipment are obtained from international markets where conditions tend to be uniform, labor costs in production become a significant differentiating factor(Pires Jr et al. 2009). Consequently, this thesis will also investigate the influence of digital twin technology on decision-making processes related to the management of labor hours within shipyard operations. This will partly be done with a case study about automating labor hours estimation for new potential projects with the help of digital twin technology.

1.2 Research Question and Objectives

The primary objective of this master thesis is to investigate how to introduce and implementing digital twin technology in Norwegian shipyards, with a special focus on assessing its potential in operational improvement and analyzing its impact on decision-making processes, particularly in the management of labor hours.

To achieve this objective, the following research questions have been formulated:
RQ1: How can the concept of a digital twin be introduced and applied in Norwegian shipyards to improve their operations?

This question aims to explore the potential applicability of digital twin technology within the context of Norwegian shipyards. It will investigate which operational areas might benefit most from implementing digital twins and how such implementation might be carried out effectively.

RQ2: How can digital twin technology influence decision-making processes in managing labor hours within shipyard operations?

This question is designed to explore into the impacts of digital twin technology on resource management within shipyard operations. It seeks to uncover how digital twins could enable more informed, efficient, and strategic decision-making about managing labor hours. In order to address the stated research question, a case study focusing on the estimation of labor hours will be systematically conducted. This empirical approach will facilitate a comprehensive exploration of the real-world implications of the research question within the targeted context.

This research aims to explore into the applicability and impact of digital twin technology in the realm of Norwegian shipyards, specifically focusing on operational optimization and decision-making processes related to labor hour management. The first research question centers on the concept of digital twin technology, seeking to identify the most beneficial areas for its application and to propose a viable roadmap for its effective integration into shipyard operations. This will involve an examination of the current state of operations, identification of areas for improvement, and the proposal of strategic applications of digital twin technology to enhance shipyard performance. The second research question ventures into the sphere of decision-making, particularly how digital twin technology can benefit labor hour management. This will encompass an investigation into the predictive and analytical capabilities of digital twin technology and how these can lead to more precise labor hour estimation, more efficient resource allocation, and more strategic decision-making. It is hoped that through addressing these research questions, the study can contribute towards the pioneering implementation of digital twin technology in Norwegian shipyards

1.3 Research Scope

The focus of this master's thesis is centered on the application of Digital Twin Technology within the context of Norwegian shipyards. The investigation will not only leverage studies directed at Norwegian shipbuilding but also seek to adapt these implementations to accommodate the unique attributes of the Norwegian shipbuilding industry.

Given that the research on digital twin technology is lacking in the shipbuilding industry, this research will additionally explore its application in other manufacturing settings. This comparison aims to glean potential insights and strategies that could be transposed to enhance the effectiveness of digital twin technology in the shipbuilding sphere

1.4 Structure

Chapter 1 Introduction	The introduction provides the background and motivation of the research project, research questions, objectives, and scope and the structure of the master thesis
Chapter 2 Methodology	The methodology presents how the project was addressed based on epistemological and philosophical considerations, describes the research methods used, and justifies the approach
Chapter 3 Theoretical background	This chapter presents the theoretical background for the research. This includes theory around digital twins, its enabling technologies, and resource and capacity management s
Chapter 4 Norwegian Shipyards	Chapter 4 presents the characteristics and the current status of operational management at the Norwegian shipyards found through the literature study and interviews with experts and representatives from case study yard.
Chapter 5 Strategic Areas Of Deployment of DT	This chapter explores potential applications of the digital twin concept in shipyards, examining benefit, drawback, and requirements drawn from literature and interviews
Chapter 5 Case study	This chapter presents a specific case of digital twin application in the case shipyard. The case study will automate labor hours estimation for new potential projects with the help of digital twin technology.
Chapter 7 Discussion	The discussion aims to answer the RQ. It provides a discussion around how the concept of digital twin can be introduced and applied in Norwegian shipyards and how it can influence decision-making in managing labor hours.
Chapter 8 Conclusion	The conclusion summarises the finding and discusses to what degree the RQ and objectives has been answered and fulfilled. The limitations of the project and further research are also provided.

2 Methodology

This chapter presents and justifies the methodology approach of this master thesis. First, it covers the theoretical methods and then the empirical methods been utilized in this study.

The methodology is depicted as a structured strategy to problem-solving(Rajasekar et al. 2006). The techniques used for collecting data and information are referred to as research methodologies(Karlsson 2010). Karlsson (2010) states that the most significant characteristic of good research is that, methodologically, it is well done.

Research methods are divided into two main categories: Quantitative and Qualitative, with some being categorized as a mix of both (Almalki 2016). Quantitative research methods concentrate on the objective reality that exists outside of observations. This is commonly done by gathering numerical data from studies, experiments, and similar techniques. Conversely, qualitative research methodology centers around deriving meaning from the viewpoints and experiences of the focus group participants. This could be achieved through literature review, observations, and interviews (Almalki 2016). This thesis predominantly utilizes a qualitative approach in its methodology. This can bring benefits such as:

- Greater depth of understanding: Qualitative research allows researchers to explore a topic in-depth, providing a rich and detailed understanding of the phenomenon being studied.
- Flexibility: Qualitative research is often more flexible than quantitative research, allowing researchers to adapt and modify their research questions and methods as the study progresses.
- Participant perspectives: Qualitative research often relies on the perspectives and experiences of participants, providing valuable insights that may not be apparent through more traditional research methods.
- Theory development: Qualitative research can be used to generate new theories or to develop existing ones further.
- Exploration of complex phenomena: Qualitative research is well-suited for exploring complex and multifaceted phenomena that may be difficult to measure quantitatively.
- Data richness: Qualitative research often generates a large amount of data, providing a rich resource for analysis and interpretation.

Overall benefits that fit the purpose and objectives of this master thieses because the focus is on exploring and not measuring as in a quantitative approach. There are also drawbacks to qualitative methods, such as subjectivity, generalizability, interpretation, validity, and ethical concerns. It's essential to have knowledge of the flaws in an attempt to minimize their effects .

However, the case study adopts a mixed-methods approach, integrating both qualitative and quantitative research methods. Quantitative methods, including statistical data analysis, descriptive statistics, exploratory data analysis, correlation analysis, predictive modeling, and model validation, are utilized to process structured numerical data and identify significant patterns and predictive variables. Concurrently, qualitative methods, such as semi-structured interviews, direct observations, and document analysis, facilitate a nuanced, in-depth investigation of digital twin implementation within the real-life context of the selected shipyard. This combined approach enables a comprehensive understanding of the research problem, leveraging the strengths of both methods (Almalki 2016).

This research has been conducted with a defined set of research questions and project objectives. A literature study looking at the topics covered in this project was conducted to answer the research question and meet the objective. In addition, semi-structured interviews with experts in the Norwegian shipbuilding industry and representatives from the case study yard are conducted to fill in gaps and get practical information. The results from the literature study, interviews and the case study have then been used as a foundation for evaluating the potential for the digital twin concept in different areas of the Norwegian shipyards.

Fundamentally, there are three methods for constructing a rational argument: deduction, induction, and abduction (Karlsson 2010). These methods vary in the order they integrate three elements: rules, observations, and outcomes. Karlsson (2010) states that principles originate from theoretical frameworks, observations stem from empirical evidence, and outcomes are produced via data analysis. The study discussed in this dissertation employs a cyclical strategy that merges existing scholarly works and empirical findings. Given the limited research in the area of interest, the emphasis has not been on verifying established theories, but rather on identifying significant characteristics that have not been previously explored in the literature. As a result, the investigation aligns more with an inductive reasoning approach than a deductive one.

2.1 Theoretical methods

The theories employed in this study is extracted from the literature study. The objective of the literature study is to compile and evaluate the current body of knowledge and perspectives related to the selected subject while examining their merits and limitations. It is essential to ascertain if the identified issue has been addressed satisfactorily, and if not, to gain an understanding of the current status of the problem. (Rajasekar et al. 2006). A significant aspect involves examine existing papers to establish authority and legitimacy to the research. This also facilitates the identification and bridging of research gaps that may limit the scope (Croom 2010) (Yin 2014). As the research process approached its conclusion, existing studies were employed to corroborate and analyze the results derived from the empirical investigation. The literature review plays a vital role in addressing research questions.

The literature study primarily aimed at establishing a foundational framework pertaining to the research questions. Additionally, it sought to enhance the understanding of the intrinsic attributes of a shipyard and the challenges associated with its operations. For the purpose of the study, two categories of search terms were devised: primary search terms and supplementary search terms. The supplementary search terms were predominantly employed to refine the breadth of the search and were amalgamated with the primary search terms as necessitated.

Primary search words	Supplementary search words
Digital twin Digital twin technologies	Manufacturing Internet of Things machine learning Engineer to order (ETO) Shipyard State-of-the-art Benefits Challenges
Shipyard Norwegian shipyard Norwegian Shipbuilding industry Engineer to order (ETO)	Challenges Characteristics Administrative processes Production processes Current status Outfitting Industry 4.0 Shipyard 4.0 Data gathering
Resource management Labor hour managment Lead time prediction Cost calculations	State-of-the-art Engineer to order (ETO) Construction Data gathering Machine learning

Table 1: Search words for Literature study

Relevant literature was mainly found by searching in the databases Google Scholar, Springer, Scopus, NTNU open, and Oria. When relevant papers were found, the abstract part was read. This was done to get an accurate impression of how suitable the report was for this thesis. The papers that were identified to be relevant were stored for further study and grouped thematically into categories.

After the first selection of papers was identified snowball sampling technique was used. Snowball sampling is a technique where the references in already found papers are used to find additional papers (Goodman 1961).

The limitation of a literature study is the relevance of previously published papers. Research with peer-reviewed articles and high-quality journals was prioritized to reduce the negative impact of this limitation. Google Scholar Metrics and the Association of Business Schools' "Academic Journal Quality Guide" were used for the journals with debatable quality (Morris et al. 2009). Considering the early development phase of digital twin applications in shipyard environments, this study also encompasses a study of digital twin deployments in related manufacturing sectors. This extended scope aims to supplement our understanding and potentially extend the relevance of the results, in light of the limited existing literature expressly concentrating on digital twin usage within shipyards.

An earlier research project delving into the introduction and application of machine learning was conducted, employing a literature study and expert interviews for data collection (Liland 2023). Portions of this information form the theoretical backbone of this master's thesis. Consequently, parts of that research paper are directly incorporated into this thesis, while others have been modified to align with the new objectives.

The references manager software BibTex stores the references throughout the project. BibTex was chosen because of its good collaboration with Overleaf. The references were mainly imported to Overleaf by the import function in google scholar.

2.2 Empirical methods

In order to thoroughly investigate the research problem and examine it from multiple perspectives, empirical methods are employed. These methods are pivotal in acquiring data and understanding the intricacies of the problem under study. The use of empirical methods, such as surveys, experiments, and observational studies, directly contributes to answering the research questions. Inspired by Voss et al. (2002), Karlsson (2010), and Yin (2014), the integration of empirical methods is deemed suitable for this study based on the following arguments:

- Empirical methods are widely used in scientific research, offering valuable insights into various phenomena and contributing to the development of new theories (Voss et al. 2002).
- The focus of this study is on the current phenomenon in real-life settings, which can be effectively investigated using empirical methods (Karlsson 2010).
- Empirical methods allow researchers to gain a comprehensive understanding of complex processes and phenomena, such as the one being investigated in this study (Yin 2014).
- The use of multiple empirical methods helps in triangulating data sources and ensuring the validity and reliability of the research findings (Denzin 1978).

Interviews, experiments, and observational studies are common empirical methods employed in scientific research (Bryman 2012). In this study, the methods are combined to gain a holistic perspective of the research problem and effectively address the research questions.

However, it is essential to acknowledge the limitations associated with these empirical methods. For instance, interviews may be subject to subjective data. The researcher's interpretation of the data may also introduce biases (Creswell and Poth 2018). Experiments can be affected by various confounding variables that may influence the results, limiting the generalizability of the findings (Campbell and Stanley 1963). Observational studies may face challenges in establishing causality and may be prone to observer biases (Denzin 1978). To overcome these limitations, multiple empirical methods, triangulated data sources, and use rigorous data analysis techniques are deployed to ensure the validity and reliability of the findings.

2.2.1 Case study

The case study methodology is a research design that allows for an in-depth investigation of a specific phenomenon within its real-life context (Yin 2014). A case study is characterized by an in-depth examination of a single instance or case, with the aim of generating insights and understanding that can be applied to a broader group of cases (Gerring 2006). While a case study itself does not directly address the research questions or objectives, it can produce valuable assumptions that, when further developed, contribute to the formation of a hypothesis on the subject (Shuttleworth 2008).

A case study methodology is particularly suited for this research as it allows for an in-depth understanding of the complex processes and interactions involved in implementing and using digital twins in shipyards (Baxter and Jack 2008). Moreover, the case study approach facilitates the examination of the specific context of the Norwegian shipbuilding industry, which may offer unique insights into the opportunities and challenges associated with digital twin technology (Flyvbjerg 2006).

For this study, a single shipyard in Norway was selected as a case to investigate the use of digital twins. The selection criteria include the shipyard's size, experience with digital twin technology, and willingness to participate in the research. The shipyard in the case study was chosen as they have collaboration with NTNU, necessary available data, and a willingness to explore how the digital twin concept can benefit them.

Multiple sources of evidence will be used to collect data for this case study to ensure the research findings are robust, reliable, and valid (Yin 2014). These sources included:

a. Semi-structured interviews: Key stakeholders, such as shipyard managers and Experts, were interviewed to gain insights into their experiences with digital twin technology (DiCicco-Bloom and Crabtree 2006).

b. Observations: Direct observations of the shipyard's operations and the use of their information systems were conducted to better understand the processes involved and identify potential challenges (Kawulich 2005).

c. Document analysis: Relevant documents, such as project reports, technical specifications, and historic data, were analyzed to gather information about the implementation and impact of digital twins in the shipyard (Bowen 2009).

Prior to data collection, informed consent was obtained from all participants, ensuring they understand the purpose of the research, their rights, and the confidentiality measures in place (Creswell and Poth 2018). All data will be anonymized and securely stored to protect the participants' privacy and maintain confidentiality (Association 2013).

The case study was divided into 7 parts:

1. Background research conducted: The case study yard, its history, and its operations were researched to comprehend the shipbuilding process specific to the company. Literature on the digital twin concept and its applications in the shipbuilding industry was also reviewed.

2. Data collection: The dataset was provided by the case company. It contained all vessels built in the shipyard with the same details as new potential projects will have attached and hours used for each department on the projects.
3. Data analyzed: The data on vessels built, hours used, and various parameters was cleaned and organized. Descriptive statistics and exploratory data analysis were performed to identify patterns and trends.
4. Relevant factors selected: The most important factors that affect the hours needed for building a vessel were identified using correlation analysis, factor importance from decision trees, and literature background.
5. The predictive model developed: With the selected factors, a predictive model was created to estimate the hours needed for new projects. Experiments were conducted with regression models, clustering, and ensemble methods like random forests.
6. Model validated: The data were divided into training and testing sets to assess the performance of the model. The model's performance was assessed using both the coefficient of determination (R-squared) and Mean Absolute Error (MAE), jointly providing a comprehensive depiction of the model's accuracy and predictive quality.
7. Digital twin concept implemented: The potential for integrating the digital twin concept into the shipyard operations was explored, and potential benefits were identified.
8. Case study report prepared: The findings were documented in a case study report. The results were then presented later with a discussion following to answer how digital twin technology can influence decision-making processes in managing labor hours within shipyard operations

Throughout the process, I actively engaged with stakeholders at the shipyard, incorporating their input and expertise to better understand the practical implications of my findings.

2.2.2 Semi-structured interview

Semi-structured interviews were conducted with experts in the Norwegian shipbuilding industry, experts in digital twin implementation, and representatives from the case study yard. The interviews were conducted through meetings, Microsoft Teams, and emails to compare the information gathered from the literature study and collect practical and missing information. This interview paradigm is particularly well-suited for undertaking exploratory, explanatory, or evaluative research. Semi-structured interviews afford researchers the flexibility to delve into various facets and identify the disadvantages and advantages associated with these diverse aspects(Yin 2014).

The decision to conduct this method was motivated by evaluation and exploratory research. Evaluate and compare literature study findings and opinions from experts and representatives from the field. Explore and discover what interview participants think is important to consider when looking into the potential of digital twins in Norwegian shipyards.

The conduction of the semi-structured interview was based on the explication of Becker et al. (2012), Matthews and Ross (2010), and Yin (2014). An interview guide was made and can be found in Appendix A. This ensures that the relevant question is discussed and keeps the conversation within the topic of interest. No strict order of questions was kept. The interviewer has the possibility to ask additional questions in response to the replies from participants. the interviews were adapted to different participants as this is one of the advantages of this research method.

Information gathered through the interviews was used to explore the Norwegian shipbuilding industry in depth and to capture the complexity and richness of the interview's experiences and perspectives. As far as possible, all of the interviews were recorded and transcribed. The summary of the interviews was then sent to the participants for verification and to be used in the final report. This ensured the accuracy and reliability of the data collected. Table 2 shows the list of interview objects.

Interview objects

Marco Semini	Expert in Norwegian shipbuilding industry
Case study representative 1	Representative from case study yard
Case study representative 2	Representative from case study yard
Employee of a major consulting firm	Representative from their digital twin division

Table 2: Interview objects

3 Theoretical background

This section serves as the foundational platform for this study, delving into a comprehensive exploration of key concepts and themes. These include the introduction to the concept of digital twins, its implications and enabling technologies, and the intersection with Industry 4.0. Resource and capacity management will be presented and the unique characteristics and challenges of the shipbuilding industry. This overview provides an essential backdrop, laying the groundwork for subsequent analyses and discussions.

3.1 Digital Twin

Although physical twins have existed for a while, the term ”Digital Twin” was first coined by Michael Grieves in 2002 during a product lifecycle management (PLM) industry presentation, defining a concept that is now commonly recognized (Kritzinger et al. 2018). The original definition of Digital Twin refers to a virtual representation of a physical system, which exists independently and is linked to the corresponding physical system. The ideal digital representation should encompass all the information about the system asset that could be obtained from its comprehensive inspection in the real world (Grieves and Vickers 2017). Gaessgen and Strahel (2012) give a more detailed and widely recognized definition: “digital twin is an integrated multi-physics, multi-scale, probabilistic simulation of a complex product and uses the best available physical models, sensor updates, etc., to mirror the life of its corresponding twin.” (Negri et al. 2017) found through a literature analysis sixteen different definitions of a digital twin. The figure underneath shows the rapidly growing number of published articles about digital twins.

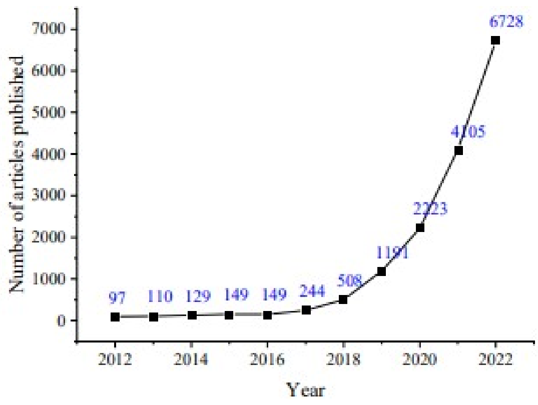


Figure 1: Graphical representation of the DT technology-related literature publication trend from 2010 to 2022 (Zhihan and Fridenfalk 2023)

One of the main benefits of digital twins is their ability to provide real-time data on the performance and behavior of their physical counterparts. By leveraging sensor data and other sources of information, digital twins can be used to monitor and predict the behavior of complex systems, enabling proactive maintenance and optimization. This has numerous applications across a wide range of industries, from manufacturing and construction to healthcare and transportation (Lim et al. 2020). In addition, digital twins can facilitate collaboration between different teams and stakeholders, providing a shared understanding of a system's behavior and enabling more informed decision-making (Tao et al. 2018).

Moreover, digital twins can be used to simulate different scenarios and test various design modifications without the need for physical prototypes, reducing the time and cost of product development (Tao et al. 2018). This is particularly valuable in industries where safety and reliability are critical. Digital twins can also support remote maintenance and repair by providing technicians with a virtual representation of the system, allowing them to diagnose and address issues more quickly and accurately. Additionally, digital twins can be used to analyze and optimize system performance over its entire lifecycle, enabling continuous improvement and cost savings. As technology continues to evolve and improve, digital twins have the potential to revolutionize the way we design, operate, and maintain complex systems, leading to increased efficiency, sustainability, and safety (Lim et al. 2020).

3.1.1 Level of integration

The concept of a Digital Twin is conventionally conceptualized as a digital replica that accurately mirrors the attributes and characteristics of a physical entity (Kritzinger et al. 2018). This definition is commonly referred to as a Digital Model, Digital Shadow, or Digital Twin, depending on the context. However, there is variation in the level of data integration between the physical object and its digital counterpart. Some digital representations are created manually and are not linked to any physical object, while others are fully integrated with real-time data exchange. In light of this variation, Kritzinger et al. (2018) suggests a classification system that categorizes Digital Twins into three subcategories based on their level of data integration.

Digital Twin
One might refer to it as Digital Twin if the data flow between an existing physical object and a digital object are fully integrated in both directions. In this type of integration, the digital object could serve as a controlling mechanism for the physical object. Additionally, there may be other objects, whether physical or digital, that cause changes in the state of the digital object. Changes in the state of the physical object would directly trigger corresponding changes in the digital object and vice versa. Creating this information flow typically requires a combination of various technologies, including data acquisition and integration, modeling and simulation, and real-time analytics. At its core, a digital twin relies on a network of sensors and other data sources to collect information about the physical system in real-time. This data is then integrated and processed using various software tools to create a digital representation of the system (Kritzinger et al. 2018).

Digital Shadow
Based on the definition of a digital Twin, if the automatic data flow from the digital object to the physical object is replaced by manual data flow, one might refer to it as a digital shadow. Changes in the state of the physical object lead to changes in the state of the digital object, but not the other way around(Kritzinger et al. 2018). A visual example can be seen in Figure 2.

Digital Model
A digital model refers to a digital portrayal of an actual or projected physical object that lacks automated data interchange between the physical and digital versions. The digital rendering may feature a more or less extensive depiction of the physical object. These models can include various representations, such as simulated models of proposed factories, mathematical models of new products, or any other models of physical objects that do not use automated data integration. Even though digital data from existing physical systems may be utilized in developing such models, all data exchange occurs manually. Any modification in the physical object's condition does not directly affect the digital object and vice versa(Kritzinger et al. 2018).

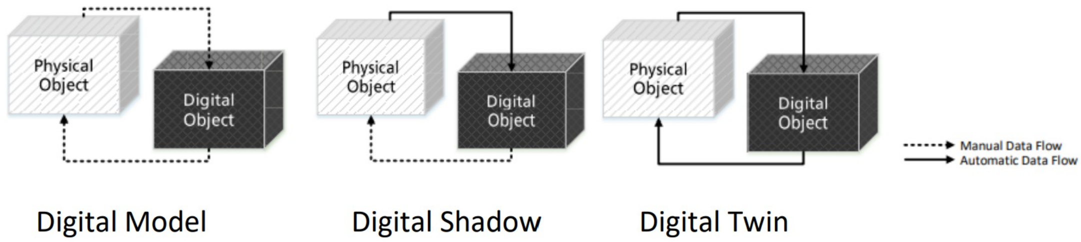


Figure 2: Data flow in Digital Model, digital Shadow, and Digital Twin (Kritzinger et al. 2018)

3.1.2 Digital twin in the manufacturing context

The concept of digital twins in manufacturing involves the utilization of virtual models for simulating and enhancing the efficiency of physical production systems. This might encompass the simulation of machine operations, assembly line processes, and the comprehensive functioning of factories. Furthermore, it also involves the predictive analysis of product behavior as they transition through various stages of the production process (Resyard 2023).

Digital twins are revolutionizing product lifecycle management (PLM), enhancing efficiency, manufacturing, service stages, and sustainability. They blend the physical and virtual product realms, providing a digital record of a product's entire lifecycle. Especially in manufacturing, digital twins simulate real-time factory operations using data collected by thousands of sensors. This innovation, made affordable by the Internet of Things, is poised to shape the future of the industry (Tao et al. 2018).

Engineers can benefit from the real-world simulation of products offered by digital twins, enabling advanced maintenance and management strategies. This technology opens up business opportunities by predicting future manufacturing processes and shifting towards proactive practices. The future of manufacturing hinges on modularity, autonomy, connectivity, and digital twins, fostering increased productivity. Modularity enhances system effectiveness, autonomy allows intelligent responses to unexpected events, and IoT connectivity closes the digitalization loop for continuous improvement (Yang et al. 2018).

Despite the proven benefits and increasing interest in Digital Twin Technology (DTT) for manufacturing, the field of manufacturing lacks a common framework for creating and implementing digital twin models (Onaji et al. 2022).

3.2 Fundamental Technologies Enabling Digital Twins

This subsection offers a theoretical background on the underlying technologies that facilitate the successful implementation of Digital Twin (DT) systems. Given the multi-faceted nature of DT, several convergent technologies are indispensable in realizing its full potential. Notably, the Internet of Things (IoT) and especially Machine Learning (ML) are most relevant for this research and will be explained in depth.

3.2.1 Internet of Things (IoT)

The Internet of Things (IoT) is an emerging paradigm that has gained significant attention in both academia and industry over the past decade (Al-Fuqaha et al. 2015). It encompasses a vast network of interconnected devices, or "things," with embedded sensors and communication capabilities, which facilitate the exchange of data and information to enable a wide range of applications and services. These devices, ranging from consumer electronics and household appliances to industrial sensors and wearable devices, are becoming increasingly prevalent, impacting various sectors such as healthcare, transportation, agriculture, and smart cities (Whitmore et al. 2015). The IoT can revolutionize how we live and work by enhancing efficiency, enabling real-time decision-making, and fostering a more connected and data-driven society(Ashton et al. 2009).

The widespread adoption of IoT technologies has been facilitated by advancements in areas such as wireless communication, sensor technology, cloud computing, and data analytics (Stankovic 2014). These innovations have allowed for the development of cost-effective and energy-efficient IoT devices that can capture, process, and transmit data in real-time. However, the rapid expansion of the IoT landscape presents numerous challenges that must be addressed to ensure its sustainability and security. These challenges include ensuring data privacy and security, handling the vast amounts of generated data, and developing scalable and energy-efficient communication protocols(Noura et al. 2019). As the IoT continues to evolve, interdisciplinary research efforts are required to explore the opportunities and tackle the challenges associated with this transformative technology, ultimately driving its successful integration into various aspects of our lives.

The role of IoT in digital twins cannot be understated, as it forms the backbone of data generation and communication for these virtual replicas (Tao et al. 2018). IoT devices, such as sensors, actuators, and communication modules, are embedded throughout the manufacturing infrastructure and equipment, collecting real-time data on various parameters, including temperature, pressure, and vibration (Qinghua and Chunjiang 2019). This data is then transmitted to a cloud-based platform, where it is processed, analyzed, and integrated into the digital twin, enabling accurate representation and simulation of the physical environment (Varghese and Tamma 2018).

IoT-enabled digital twins can facilitate enhanced operational efficiency, predictive maintenance, and real-time decision-making, resulting in significant cost savings and reduced downtime (Lu et al. 2017). Additionally, digital twins can support remote monitoring, control, and optimization of shipyard operations, as well as improve the design and manufacturing processes through data-driven insights (Grieves and Vickers 2014). However, realizing the full potential of IoT-driven digital twins requires addressing the aforementioned challenges related to data privacy, security, and scalability, as well as developing robust data fusion algorithms and efficient communication protocols tailored for specific manufacturing environments (Si et al. 2019). By overcoming these obstacles and leveraging the capabilities of IoT, digital twins can revolutionize the shipbuilding industry and pave the way for more sustainable, efficient, and data-driven shipyards.

3.2.2 Artificial Intelligence and Machine Learning

Artificial intelligence (AI) is a rapidly evolving field of computer science that aims to develop machines and systems capable of performing tasks that would typically require human intelligence, such as learning, problem-solving, reasoning, perception, and natural language understanding (Russell and Norvig 2020). By leveraging vast amounts of data and advanced algorithms, AI systems can analyze complex patterns, make predictions, and adapt their behavior to achieve specific goals (Goodfellow et al. 2016). As a result, AI has become an essential component of various industries and applications, revolutionizing the way we interact with technology and transforming our daily lives (Brynjolfsson and McAfee 2017).

In the context of digital twins, AI plays a crucial role in enhancing the accuracy, efficiency, and effectiveness of the virtual models (Tao et al. 2018). By analyzing the continuous stream of data gathered from IoT devices, AI algorithms can detect anomalies, predict future outcomes, and optimize system performance (Shi et al. 2016). Moreover, AI-driven automation helps reduce human intervention, enabling timely decision-making and minimizing errors (Monostori et al. 2016). With the ability to learn from the data and adapt over time, AI-powered digital twins become increasingly accurate and reliable, leading to improved operational efficiency, reduced downtime, and better resource utilization (Negri et al. 2017). Overall, the integration of AI within digital twin technology allows businesses to harness the full potential of their physical assets and make more informed decisions in real time (Uhlemann et al. 2017).

There are multiple branches within artificial intelligence, with Machine learning being one of the most significant ones and the moste relevant for this paper (El Naqa and M. J. Murphy 2015). Machine learning algorithms build a model based on batches of sample data, known as training data. The model can then make decisions or predictions without being explicitly programmed to do so in code. This is achieved by supplying the algorithm with a set of data, known as the training set, which includes both inputs and, for certain algorithms, the desired outputs.

Liland (2023) discussed the potential of machine learning at Norwegian shipyards. A framework is developed and many promising applications such as demand forecast, lead time prediction, cost estimation, and process estimation are laid forward. The main implementation requirements are a Sufficient amount of quality data and other data-gathering technologies.

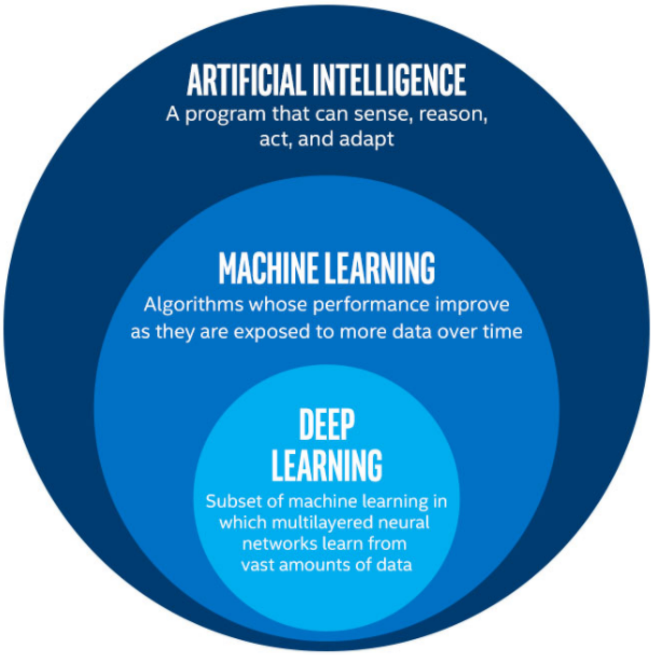


Figure 3: Relationship of Artificial intelligence, machine learning and deep learning (Robins 2017)

Supervised learning

Supervised learning is one of the most simple forms of machine learning. A supervised machine learning algorithm aims to become a trained model which transforms the input into the fitting output. (Muhammad and Yan 2015). This is done through a training set with inputs and desired output. More formally, each training example can be written as (x,y). x_i as the observed feature vector and y_i as the observed label. It's assumed that the sample (x_i,y_i) are independent and identically distributed from a certain unknown distribution. For instance, x_i may be the characteristics of a customer at a website, such as age, gender, and even more specific things; $y_i = 1$ if the person clicks on the advertisement on the site, and $y_i = 0$ if not. A supervised learning algorithm will take in the given (x_i,y_i) set and make a model for the prediction of the future. This model can then decide what advertisement should be used for a specific customer or calculate how many clicks a certain advertisement will get. While going through the training data, it will adjust itself after how close the model's output is to the desired output. The model can often perform a good estimation with the right amount of data.

In Operation management literature, the goal of supervised learning is either prediction or inference (Bastani et al. 2022) . For prediction, researchers use machine learning models to forecast variables that are later used in operational decisions. For these models, the accuracy of prediction is the measurement of the quality of the model. On the other hand, in inference problems, the interesting output is the underlining reason for how an outcome variable is generated as a function of the input data.

Regression analysis

Regression analysis is a popular machine learning application used in various scientific disciplines. It is a supervised learning algorithm that aims to predict a continuous output variable based on one or more input variables (Hastie et al. 2009). The primary goal of regression analysis is to identify the relationship between the dependent variable and the independent variables and make accurate predictions on new data.

Linear regression is the most commonly used regression analysis technique, and it assumes a linear relationship between the independent variables and the dependent variable (Draper and Smith 1998). Linear regression is simple to interpret and efficient to compute. However, if the relationship between the variables is nonlinear, linear regression may not be appropriate. In this case, non-linear regression can be used to model the relationship between the variables. Non-linear regression can take different forms such as polynomial regression, logistic regression, and exponential regression (Chai and Lin 2021). Non-linear regression is more flexible and can fit a wide range of data patterns. Therefore, the choice between linear and non-linear regression depends on the nature of the data.

The regression model is developed based on a training dataset, where the input variables are used to predict the output variable using statistical methods. The model is then used to make predictions on new data by applying the same statistical methods to the input variables. The performance of the model is evaluated using various metrics, such as mean squared error or R-squared, which measure the accuracy of the predicted output compared to the actual output.

Random forest

Random forest is a popular machine learning application that belongs to the family of decision tree-based ensemble methods. It is a supervised learning algorithm that is used for both classification and regression problems (Breiman 2001). The primary goal of random forest is to create a predictive model by combining the outputs of multiple decision trees, each trained on a random subset of the training data. The final prediction is based on the average or mode of the predictions from individual trees, which helps to reduce overfitting and improve the accuracy of the model.

Random forest is widely used in various scientific disciplines for a range of applications, such as predicting customer churn, detecting fraudulent activities, diagnosing diseases, and predicting lead time (Cutler et al. 2007). It is particularly useful for handling high-dimensional data with complex nonlinear relationships between the input and output variables. Random forest is also robust to missing data and noisy features, making it a popular choice for many practical applications.

Random forest works by creating multiple decision trees using a bootstrapped sample of the training data and selecting a random subset of features at each split. The trees are grown independently and then combined to make predictions on new data. The final prediction is obtained by averaging or voting over the predictions from individual trees. The algorithm also provides a measure of feature importance, which can help to identify the most relevant features for the prediction task (Liaw and Wiener 2002). By combining the outputs of multiple decision trees, the random forest can produce a more accurate and stable model than a single decision tree.

In the context of a Random Forest model, Gini importance, also known as Mean Decrease Impurity, is a widely recognized metric for feature selection. It measures the total reduction of the criterion, often impurity, brought by a feature through its inclusion in the trees of the forest. Essentially, the Gini importance of a variable assesses the degree to which that variable contributes to the homogeneity of the nodes and leaves in the resulting Random Forest. Higher Gini importance indicates a more significant contribution of the variable in the model's prediction, making it a critical aspect in the model's performance Strobl et al. 2007.

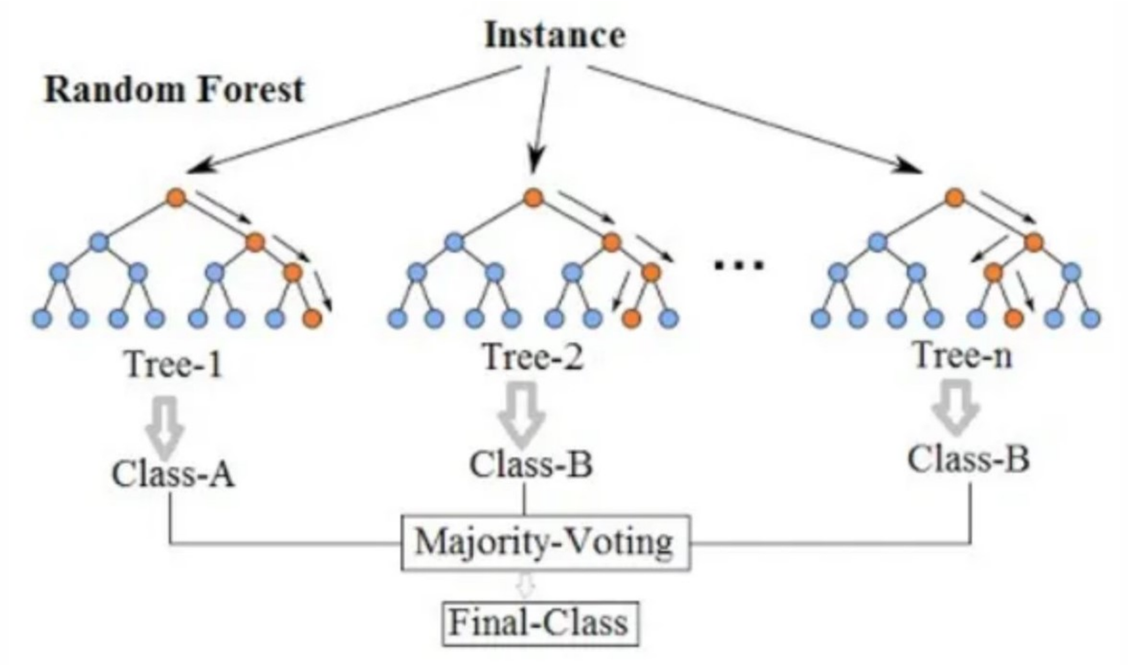


Figure 4: Random Forest example (Inc. n.d.)

Unsupervised learning

Unsupervised learning algorithms often have an essential role in structuring the data for predictions (Hinton and Sejnowski 1999). One can customize the prediction model for each cluster to achieve better accuracy. Studies show that predictions based on clustering have higher accuracy than the implementation of predictions. In operational management literature, one often uses latent variable models for prediction problems (Bastani et al. 2022).

Clustering

Clustering is a popular machine learning application that is widely used in various scientific fields such as image analysis, social network analysis, and bioinformatics. It is an unsupervised learning technique that involves grouping similar objects into clusters based on their features or characteristics (Jain et al. 2010). Clustering is used to discover hidden patterns or structures in the data, and it can be helpful in identifying outliers, detecting anomalies, and reducing the dimensionality of the data.

Clustering works by assigning each data point to a cluster based on the similarity of its features with the other data points in the same cluster. The objective of clustering is to maximize the similarity within each cluster and minimize the similarity between different clusters. Clustering algorithms can be broadly classified into two categories: hierarchical clustering and partitional clustering. Hierarchical clustering methods group data points into a tree-like structure based on their similarity, while partitional clustering methods divide the data into non-overlapping groups or partitions (Han et al. 2011).

Clustering algorithms can also differ in terms of the similarity measure used, the number of clusters, and the algorithm's optimization criteria. Some common clustering algorithms include k-means, hierarchical clustering, and density-based clustering. The choice of the clustering algorithm depends on the nature of the data, the objectives of the analysis, and the available computational resources (Jain et al. 2010).

KMeans

K-means is a widely used unsupervised clustering algorithm, introduced by MacQueen in 1967 MacQueen 1967b. The algorithm aims to partition a dataset into K distinct, non-overlapping clusters based on the similarity of data points within each cluster. K-means operates iteratively, assigning data points to their nearest cluster centroids and updating the centroids based on the mean of the assigned data points until convergence is achieved or a predefined stopping criterion is met. The algorithm is widely used in various fields, including image segmentation, anomaly detection, and market segmentation, owing to its simplicity, efficiency, and ease of implementation (Jain et al. 2010).

The K-means algorithm begins by initializing K cluster centroids, either randomly or using an initialization method such as K-means++ (Arthur and Vassilvitskii 2007). The algorithm then iteratively refines the centroids through a two-step process: assignment and update. In the assignment step, each data point is assigned to the nearest centroid based on a distance metric, typically Euclidean distance (MacQueen 1967b). In the update step, the centroids are recalculated as the mean of all data points assigned to each cluster. These two steps are repeated until the centroids' positions stabilize, or a predefined stopping criterion is met, such as a maximum number of iterations or a minimum change threshold in the centroids' positions (Jain et al. 2010). The final clustering result is determined by the assignment of data points to the nearest centroid in the last iteration.

Despite its popularity, K-means has a few limitations. One primary concern is the need to specify the number of clusters (K) in advance, which is not always evident for a given dataset. Several techniques, such as the elbow method and silhouette analysis, have been proposed to address this issue (Jain et al. 2010). Another limitation is the algorithm's sensitivity to the initial placement of centroids, which can lead to convergence to local optima. To overcome this, researchers have proposed methods like the K-means++ initialization (Arthur and Vassilvitskii 2007).

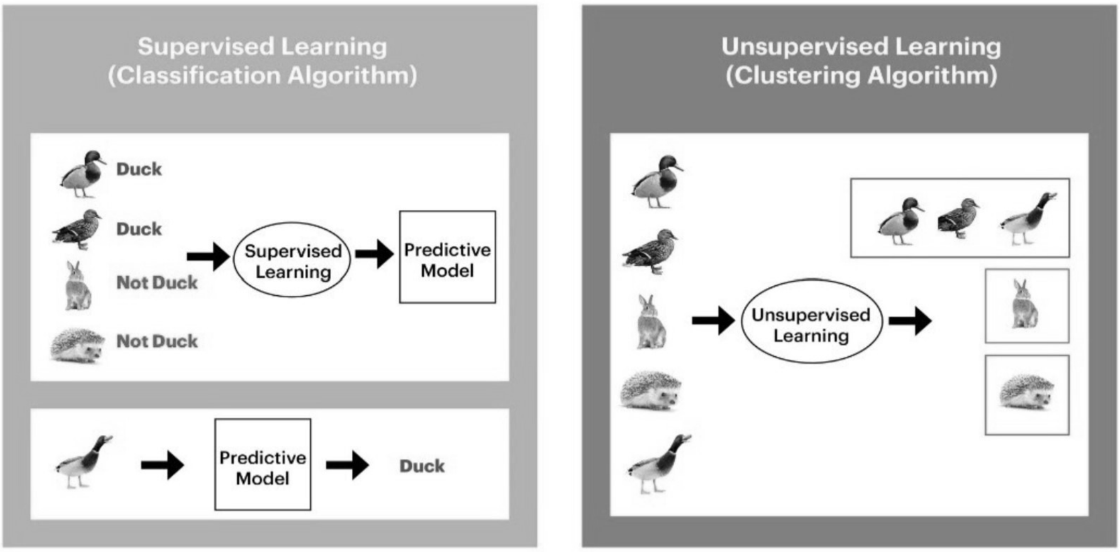


Figure 5: Supervised learning vs Unsupervised learning (Amiri et al. 2018)

Reinforcement learning

Reinforcement learning is also a key branch of machine learning algorithms. The main trait of reinforcement in compassing to supervised learning is that you don't need to label your input (Kaelbling et al. 1996). The focus of the algorithms is on finding a balance between exploration (of uncharted territory) and exploitation (of current knowledge). This means that a reinforcement algorithm will explore uncharted territory until it finds the best possible solution. Reinforcement learning can be an effective tool for improving the efficiency and effectiveness of operational management systems in a variety of industries (Bastani et al. 2022). In operational management, it's often used to optimize decision-making processes and control systems.

In reinforcement learning, an agent learns to make decisions by interacting with an environment. In this paradigm, an agent takes actions within an environment, and for each action, it receives feedback in the form of rewards or penalties. The agent's objective is to learn a policy, which is a strategy for choosing actions, that maximizes the cumulative reward over time, often by exploring the possible states and actions and learning to favor those that result in more favorable outcomes (Kaelbling et al. 1996).

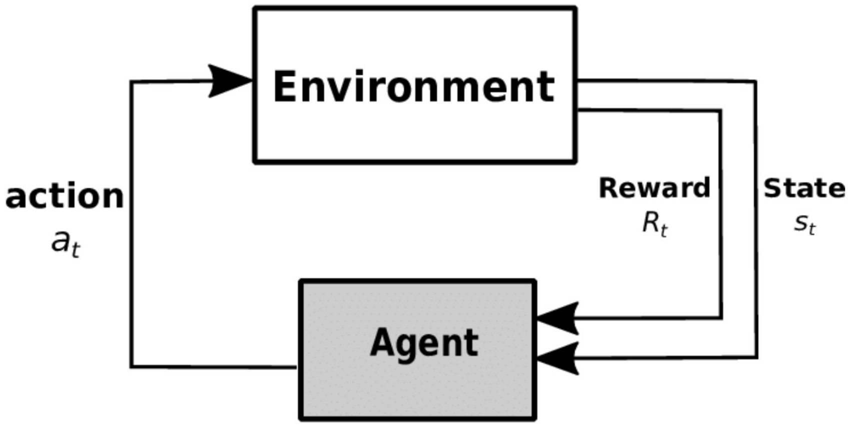


Figure 6: Reinforcement learning example (Amiri et al. 2018)

Deep learning
Deep learning is a subsection of machine learning as shown in figure 3. Deep learning uses multiple layers to progressively extract higher-level features from the raw input (Deng, Yu et al. 2014). For instance, in Digital image processing, the first layers will identify edges, and the later layers will identify more letters, digits, and concepts more relevant to humans. In the last ten years, the best-performing ML systems have resulted from deep learning, for instance, speech recognition on phones and Google automatic translator. Deep learning is a new name for a Machine learning technique that has been going in and out of fashion for over 70 years, Artificial Neural Network.

Artificial neural Network
Artificial neural networks (ANNs), usually called neural networks, is a computing system inspired by the biological neural networks that constitute a real brain (Brahme 2014). A Neural network tries to mimic how biological neurons signal to one another.

An ANN consists of layers of nodes or artificial neurons (Brahme 2014). ANNs contain an input layer, one or more hidden layers, and an output layer. Every node connects to another and has a connected weight and threshold. The node is activated when the output is over the threshold and will then send data to the next layer in the network. If the output is under the threshold, no data is passed along from that node.

As a machine learning technique, ANNs rely on training data to both learn and improve their accuracy (Schmidhuber 2015). A fine-tunes neural network is looked at as one of the most powerful tools in machine learning. Tasks such as image recognition can take minutes versus hours when compared to manual identification by an experienced worker. This can be used in many phases of the shipping industry, for instance, it can be a potential way to do quality control to minimize manual work.

A challenge with an ANN is that it works like a black box (Dayhoff and DeLeo 2001). It gives input and gives back calculated output, but the calculations are hidden. This makes neural networks not optimal when reasoning around a decision is important.

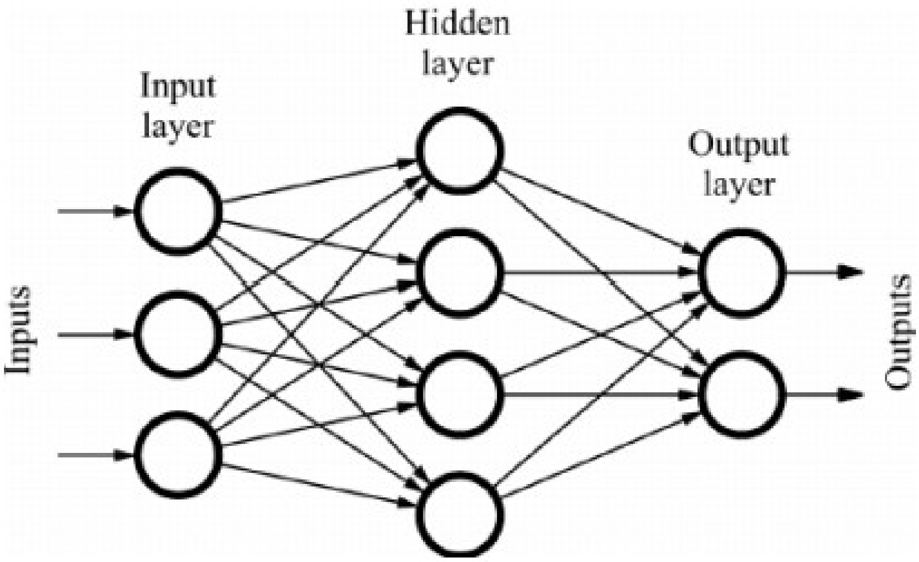


Figure 7: Artificial neural network (Stephansen-Smith 2020).

Data quality and cleaning

Data quality is a multifaceted concept that is challenging to define due to the diverse contexts in which data is used, as well as the different perspectives of end users, data producers, and data custodians (Fürber 2016). For instance, data quality will be different from consumer and business perspectives. From a business perspective, data quality is data that are fit for their intended uses in operations, decision-making, and planning (Fleckenstein et al. 2018).

Data quality expectations, specifications, and requirements are stated in terms of the characteristics or dimensions of the data, such as (Taleb et al. 2018)(Mahanti 2019):

- Accessibility or availability
- Accuracy or correctness
- Comparability
- Completeness or comprehensiveness
- Consistency, coherence, or clarity
- Credibility, reliability, or reputation
- Flexibility
- Relevance, pertinence, or usefulness
- Timeliness or latency
- Uniqueness
- Validity or reasonableness

Data cleaning refers to the process of identifying and correcting or removing errors, inconsistencies, or inaccuracies in a dataset (Wu 2013). This is important because incorrect data can lead to unreliable results and algorithms. While the specific steps involved in data cleaning may vary depending on the specific characteristics and needs of a dataset, it is essential to have a consistent and structured approach to ensure the integrity and reliability of the data. This may involve identifying and correcting errors, eliminating duplicates, formatting data consistently, and verifying the accuracy and completeness of the data.

Insufficient data handling

When applying machine learning there is often a question of how much data is needed(Bzdok et al. 2017). There is no single right answer to how many samples are needed to reach the accepted prediction performance.

The main consideration of machine learning is the n-p ratio (Lever et al. 2016). In this case, the n represents several samples and the p represents several variables per observation. Machine learning is usually most effective in the high-dimensional setting, that is when p is significantly higher than n, with over hundreds of variables to be fitted.

The complexity of the machine learning algorithm is also one of the primary considerations to be looked at and should be calibrated with the complexity of the given data (Bzdok et al. 2017). More data is needed for the more complex and sophisticated algorithms. For instance, a deep neural network is made to capture complex trends and therefore needs enough data to avoid overfitting. On the other hand, simple algorithms require less data. The complicated interaction between the variables will in these cases be abandoned. Simple algorithms that do not overfit, can in some cases also outperform complex algorithms when trained with big data sets (Steck 2019).

Problems with Insufficient data

Since machine learning depends on certain amounts of data, limited data is often a problem when introducing machine learning. The problems can be arranged into different categories. This subsection will introduce the categories that are relevant to this project.

Overfitting:

Overfitting in a machine learning algorithm represents one of the prevalent pitfalls, particularly when dealing with a constrained dataset. In mathematical modeling, overfitting occurs when the model is tailored excessively to the specific dataset, adhering too closely to its nuances. As a consequence, the model may perform poorly in predicting new, unseen observations. Essentially, an overfitted model encompasses a higher number of parameters than can be justified by the data, capturing noise rather than the underlying trend (Everitt and Skrondal 2010). Figure 10 shows a visual example.

Underfitting:

A machine learning algorithm is said to be underfitting when it fails to capture the fundamental patterns inherent in the data. Analogous to overfitting, underfitting may yield satisfactory results on the training data, but the model's performance deteriorates when evaluated on testing data. This is attributed to the model's overly simplistic nature, which renders it inadequate for making reliable predictions on novel data (Everitt and Skrondal 2010).

Non- Representative data:

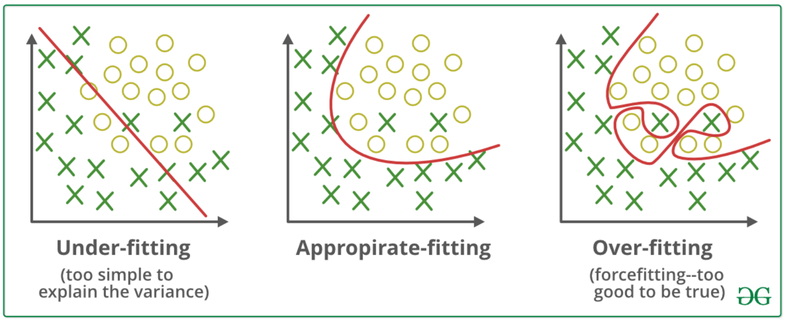


Figure 8: Example of overfitting and underfitting (Nautiyal 2019)

It is imperative for the training data to be an accurate representation to create a good machine learning model(Mannino et al. 2009). When the dataset is limited in size, sampling noise may arise if the data is non-representative, signifying that certain observations deviate significantly from the remainder of the dataset. This phenomenon can also manifest in the context of extensive datasets, where it is referred to as sampling bias (Hellström et al. 2020). The occurrence of sampling bias implies that the dataset lacks the representativeness necessary for efficacious model training.

Potential solutions for insufficient data

Model complexity:

It is important to note that adding complexity to a model does not always improve its fit or lead to more accurate predictions (Verhagen 2021). It may be necessary to carefully consider the trade-off between model complexity and fit, as well as the potential for overfitting. Steck (2019) illustrates in studies about recommender systems that less complexity can outperform more complex models.

Transfer learning:

Transfer learning constitutes a methodology in machine learning that emphasizes the retention of knowledge acquired during the resolution of a distinct problem, and capitalizing on this knowledge for the addressal of analogous challenges (West et al. 2007). For example, knowledge gained while learning to recognize big Ships could apply when trying to recognize small boats.

Data augmentation:

Data augmentation in an analysis of data is a technique used to increase the amount of data available in the training set by slightly modifying copies of already existing data or newly synthetic data from existing data (Shorten and Khoshgoftaar 2019). The augmentation of data can, for instance, be to crop, crop, or zoom an image to add more data. It acts as a regularizer and helps reduce the chance of overfitting by adding more data to the training. This technique is mostly used for image classification, speech recognition, and signal processing.

Cross validation:

Cross-validation is a widely-used resampling procedure in machine learning that aids in assessing the performance of predictive models. Its primary goal is to ensure the model's robustness and to prevent overfitting, a scenario where the model excessively adapts to the training data, hence failing to generalize to unseen data. The most common form of cross-validation, the k-fold cross-validation, partitions the original dataset into 'k' equal-sized subsets. For each iteration, one subset is used as a validation set, while the remaining subsets form the training set. The model is then trained and evaluated k times, each time with a different validation subset. The final model performance is calculated by averaging the evaluation metric across all iterations, thereby providing a more comprehensive measure of model accuracy and robustness (developers 2021).

In the realm of machine learning, cross-validation is instrumental for model selection, hyperparameter tuning, and understanding the model's variability. By providing multiple 'folds' or subsets of the dataset, it allows the model to learn from different data partitions, effectively reducing the bias caused by a single random data split. Moreover, cross-validation aids in hyperparameter tuning by identifying the best set of hyperparameters that minimize the validation error, thus optimizing the model's predictive ability (developers 2021). See figure 9 for an visual example.

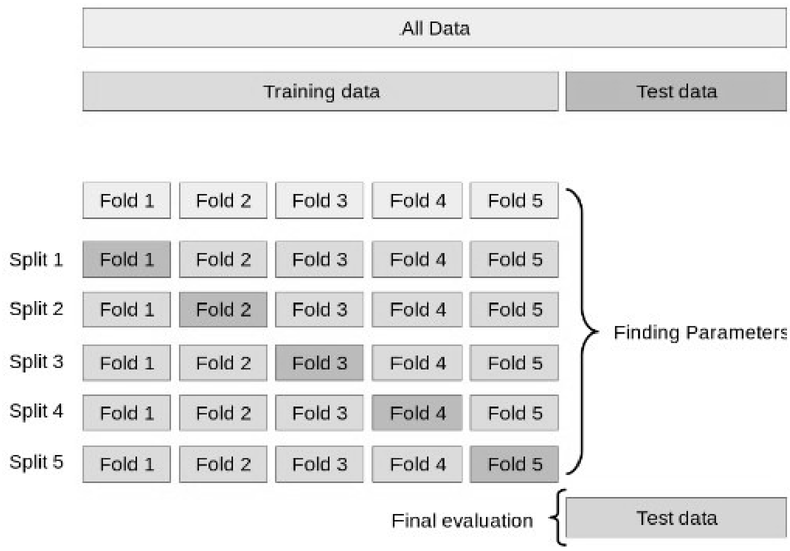


Figure 9: Train-test-split visualization (developers 2021)

The methodology for developing machine learning models

In machine learning, the main objective is the construction of models that possess the capacity to generalize effectively to novel, unseen data. To accomplish this, it is imperative to partition the available dataset into distinct subsets for training and testing purposes (Goodfellow et al. 2016). The training set is used to train the model, allowing it to learn patterns and relationships within the data. In contrast, the testing set is used to evaluate the model's performance on unseen data, providing an estimate of its ability to generalize to unseen examples. This separation helps to prevent overfitting, a phenomenon where the model becomes too specialized to the training data and performs poorly on unseen data (Hawkins 2004). The choice of the splitting ratio between training and testing sets depends on the problem and data size, with common choices being 70:30, 80:20, or 90:10.

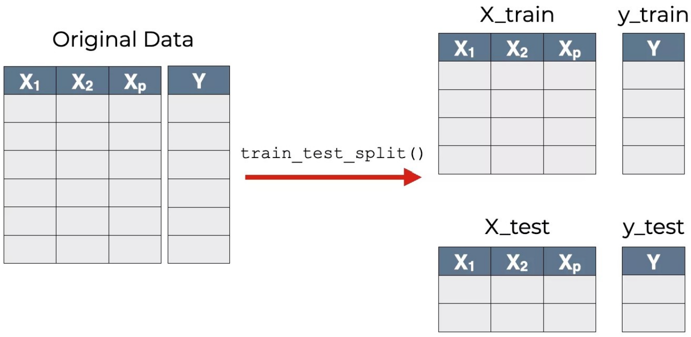


Figure 10: Train-test-split visualization (Labs 2018)

Loss functions play a critical role in training machine learning models, as they quantify the discrepancy between the model's predictions and the true values (Bishop 2006). By minimizing the loss function, the model adjusts its parameters to better fit the data. There are various types of loss functions, each suited to different tasks and data types. For example, mean squared error, mean absolute error, and R^2 are commonly used in regression tasks, where the goal is to predict a continuous value. On the other hand, categorical cross-entropy is widely used for classification tasks, where the goal is to predict which class a given instance belongs to. The choice of the appropriate loss function is crucial, as it directly impacts the model's performance and convergence during training (Goodfellow et al. 2016).

The coefficient of determination, known as R-squared (R^2), is a statistical metric that indicates the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. It is calculated using the formula

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}},$$

, where SS_{res} is the sum of squares of the residual errors, and SS_{tot} is the total sum of squares, a measure of the variance in the observed data (Miles 2005).

The Mean Absolute Error (MAE) is a statistical measure used to quantify the accuracy of predictions in a regression model. It represents the average magnitude of the errors in a set of predictions, without considering their direction. Essentially, it is the average absolute difference between observed and predicted outcomes. MAE is calculated using the formula $MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$, where y_i is the observed value, $\hat{y}_i$ is the predicted value, and n is the total number of data points. Because MAE uses absolute values, it cannot indicate the underperformance or overperformance of the model — only the magnitude of the error.

The creation of a machine learning model typically involves multiple stages. This underscores the fact that developing a Machine learning model extends beyond the model alone. Various sources outline different steps involved in the process. This analysis synthesizes the definitions provided by Chapman et al. (2000) and Géron (2022):

1. *Business Understanding:* The initial stage involves comprehending the specific business context and formulating the problem. This facilitates the selection of appropriate input features, the optimal model, and the validation of outcomes. As a critical determinant of successful results, this stage may demand the most substantial investment of time.

2. *Collect and analyze the data:* The second stage encompasses the gathering and examination of relevant data. This phase is vital for identifying patterns, trends, and correlations that may inform the development of the machine learning model.

3. *Process data:* The third stage involves preprocessing and cleaning the collected data. This step is essential to ensuring the quality and accuracy of the data, which will in turn affect the performance of the ML model.

4. *Model development:* During the fourth stage, the machine learning model is created based on the preprocessed data. This involves selecting an appropriate algorithm, training the model, and tuning its parameters to optimize performance.

5. *model validation:* The fifth stage focuses on evaluating the developed model's performance. By comparing its predictions to actual outcomes, this stage helps to identify potential issues, refine the model, and assess its generalizability to new data.

6. *Deploy:* The final stage entails deploying the validated machine learning model in a production environment. This step involves integrating the model into existing systems, monitoring its performance, and updating it as needed to maintain optimal results.

3.2.3 Other technologies

IoT and machine learning are the most relevant technologies for this study. This sub-subsection explores other vital technologies that underpin the implementation and effectiveness of Digital Twin systems. We look into other enabling digital twin technologies such as cloud and edge computing, big data analytics, augmented and virtual reality, and blockchain. These technologies, although not exhaustively detailed in this thesis, play a significant role in shaping the landscape of DT, impacting its scalability, real-time data analysis, security, interactivity, and more. Their inclusion aids in painting a comprehensive picture of the technological ecosystem surrounding DT.

Cloud computing
Cloud computing facilitates scalable storage and processing power for DTs, enabling them to handle vast amounts of data from different sources. Meanwhile, edge computing brings data processing closer to the source, reducing latency and allowing for real-time data analysis critical to DT applications (Premsankar et al. 2018).

Big Data Analytics
Big data analytics enables the processing and interpretation of large volumes of data generated by DTs, turning it into actionable insights (Qi and Tao 2018).

Augmented and virtual reality
AR and VR can provide immersive visualization and interaction with DTs, enhancing understanding and decision-making (Billinghurst and Duenser 2018).

Blockchain
Blockchain can enhance the security and trustworthiness of DT data, ensuring its integrity and traceability (Casino et al. 2019).

3.2.4 The Intersection of Industry 4.0 and Digital Twin Technology

The Fundamental Technologies Enabling Digital Twins, as outlined in the last sub subchapters, encompass various technologies intrinsic to Industry 4.0. There is an intrinsic linkage between Industry 4.0 and Digital Twins, where Digital Twins serve as a critical component in realizing the intelligent and interconnected systems that define Industry 4.0 (Pires et al. 2019).

A German government program published a report in 2011 describing Industry 4.0 as the fourth industrial revolution (Manikas et al. 2020). The report pointed towards industry 4.0, enabled by the introduction of the Internet of Things and cyber-physical systems to form smart factories - grouped under the term digitization. Industry 4.0 and digitization are often used interchangeably. Digitization relates to most aspects of society, while Industry 4.0 refers specifically to the digitization of industry. The foundational technologies of Industry 4.0, also known as the fourth industrial revolution, form the bedrock of a concept known as the digital twin.

There are different ways to conceptualize Industry 4.0. Frank et al. (2019) separate it into front-end and base technologies, as seen in figure 11. Front-end Technologies consists of an intelligent supply chain, smart working, smart manufacturing, and innovative product. Base technologies are considered to be the Internet of Things, the cloud, big data, and analytics. Machine learning will mainly be found in analytics in conceptualizing Industry 4.0.

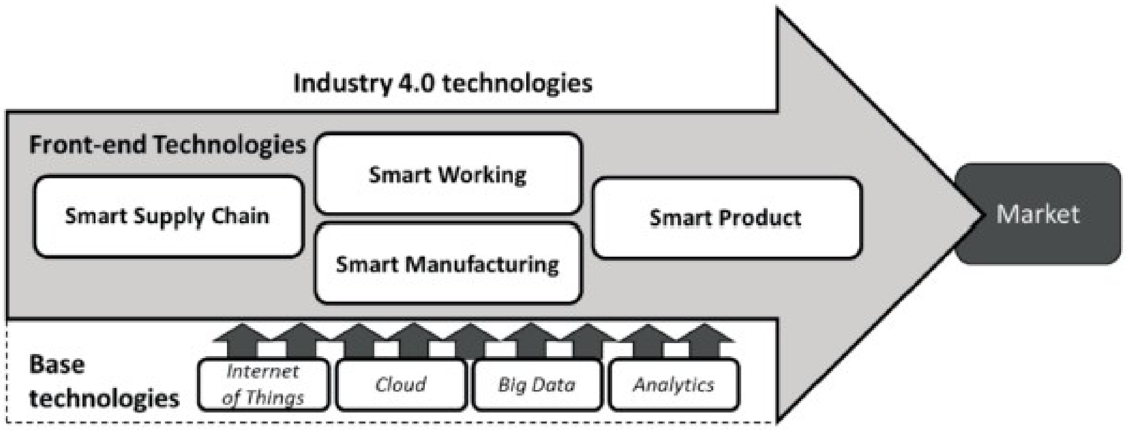


Figure 11: Conceptual framework for Industry 4.0 technologies (Frank et al. 2019)

There are different ways to group the fundamental technologies associated with Industry 4.0. Winkelhaus and Grosse (2020) group the technologies as (a) technologies that generate data, (b) technologies that handle data, and (c) technologies that use data. Machine learning is primarily found in analytics, a critical component in the functioning of digital twins. By enabling the processing and understanding of large volumes of data in real time, machine learning significantly enhances the accuracy and utility of digital twin models. Strandhagen (2022) provides an overview of one possible way of listing the technologies with descriptions.

Technology group	Description
Additive manufacturing	3D printing of objects layer by layer, based on 3D models or CAD files of the objects.
Autonomous robots	Automatic guided vehicles (AGVs), autonomous mobile robots (AMRs), and collaborative robots (COBOTS) for material handling and performing logistics operations.
Cloud manufacturing	Cloud-based solutions for sharing and exchange of data between systems, sites, and companies.
Cyber security	The secure and reliable protection of industrial production systems from cyber threats.
Data analytics	Transforming data into knowledge and actions within a manufacturing system. Big Data for analysis of large sets of real-time data, artificial intelligence, machine learning, and advanced simulations are all part of this group.
Integration of IT systems	Horizontal and vertical integration of IT systems for production management (PLM, ERP, MES).
Internet of Things	Objects equipped with sensors and actuators, enabling storing and exchange of information through network technology.
Visual technology	The visual representation of an object, in the form of augmented reality (AR), through superimposing a computer-generated 3D image in the real world, creating a virtual reality (VR), or projecting 3D images as holograms.

Figure 12: Overview and description of digital technologies in manufacturing logistics(J. W. Strandhagen et al. 2019)

Figure 12 describes machine learning as the data analytics group that transforms data into knowledge and actions within a manufacturing system. However, the table does not show the link between the technologies. For instance, predictive maintenance data collection needs IoT to gather real-time data. This data need to be analyzed through, for instance, machine learning.

Context dependence is an essential factor to consider when applying Industry 4.0 technologies (Strandhagen 2022). While it is reasonable to assume that some of these technologies can be universally applicable, some are more suitable for particular manufacturing environments. Furthermore, it appears that the effectiveness of the technologies may vary depending on the context. Consequently, it is necessary to consider the specific context when applying Industry 4.0 technologies.

The applicability of Industry 4.0 technologies in manufacturing environments may be influenced by the degree of repetitiveness (Strandhagen 2022). Strandhagen (2022) tested this hypothesis in a study that included a shipbuilding company, which had the lowest repetitiveness and the least applicability of Industry 4.0 technologies. It should be noted that this study was conducted in 2016 when Industry 4.0 was still a relatively new concept. Further research is needed to confirm and expand upon these findings.

The rise of digital twins represents a crucial development in the world of Industry 4.0. As we continue to explore and understand the optimal ways to apply and integrate these advanced technologies into various manufacturing contexts, the digital twin framework offers a compelling blueprint for how we might better leverage IoT, data analytics, machine learning, and other Industry 4.0 technologies to revolutionize manufacturing and logistics practices (Pires et al. 2019).

3.3 Resource and capacity Management

Resource Management is an integral aspect of effective planning and execution in various contexts, including project management, operations management, and organizational strategy (Kerzner 2017; Slack et al. 2018). These concepts facilitate the optimal allocation of resources and capabilities, ensuring that organizations can successfully achieve their objectives and goals.

Resource Management
Resource management entails the systematic approach of formulating strategies, orchestrating schedules, and distributing human capital, financial assets, and technological resources within a project or program framework. This process is fundamentally aimed at optimizing the allocation of resources to maximize the value derived by the organization. Effective resource management is characterized by the judicious alignment of resources, ensuring their availability in a timely manner and appropriating them to tasks that best serve the organizational objectives (Kerzner 2017). Resource management encompasses:

- Identifying and prioritizing resource requirements based on organizational goals and objectives.
- Allocating resources to different projects or tasks according to priority, availability, and constraints.
- Monitoring resource utilization to verify that resources are being used effectively and efficiently.
- Adjusting resource allocation and schedules as needed to accommodate changes in priorities, resource availability, or project requirements.

Effective resource management assists organizations in preventing resource conflicts, delays, and cost overruns while maximizing productivity and ensuring the successful completion of projects and operations (Muller 2017).

Capacity Management
Capacity management pertains to the systematic planning, monitoring, and optimization of resources to address both current and future demands (Bourne et al. 2018). This concept is critical in evaluating the ability of a system or resource, including personnel, equipment, facilities, or technology, to produce output or perform tasks. Capacity management entails:

- Assessing the existing capacity of resources.
- Identifying capacity-related gaps or bottlenecks.
- Forecasting future capacity requirements in relation to business growth, seasonal variations, and other influential factors.

- Implementing strategies to adjust capacity, such as hiring additional staff, procuring new equipment, or outsourcing tasks.

Efficient capacity management contributes to optimized resource utilization, cost reduction, enhanced service quality, and the facilitation of long-term growth (Bicheno and Holweg 2016).

In conclusion, capacity management centers on optimizing the use of available resources to meet current and future demands, while resource management focuses on planning, allocating, and controlling resources to achieve organizational goals and objectives. Both concepts play a crucial role in enabling organizations to operate efficiently, minimize costs, and support growth (Slack et al. 2018; Kerzner 2017).

3.4 Shipbuilding industry

Although there are variances in features in shipbuilding across sectors and countries, there are still some basic characteristics that are typical for the industry of shipbuilding (Dugnas and Oterhals 2008)(Lamb 2004). part of the variance is also due to the level of customization and complexity of the vessel built. Ships are either built when a customer order is received, or engineering or customization starts with the received order. Shipbuilding can therefore be classified as an engineer-to-order (ETO) or make-to-order (MTO). In Norway, most shipyards are set up as ETO companies (Hagen and Erikstad 2014).

Demand variability

Predicting short-term demand in the commercial shipbuilding industry is challenging due to its significant variability. It's influenced by market conditions, the global financial situation and also many other factors. While short-term fluctuations are moderately predictable, the medium to long-term variations can be substantial and more difficult to foresee. There is a substantial global demand for commercial vessels, with large shipyards that are cost-efficient being able to manufacture up to 40 ships annually. The industry must exhibit flexibility in scaling their capacities up or down in response to these demand fluctuations. Notably, there has been a decline in demand in recent years owing to the pandemic, and it is reasonable to anticipate a surge in demand as the industry seeks to compensate for reduced production during this period(Lamb 2004).

One-of-a-kind production

The majority of shipyards operate on an Engineer-to-Order (ETO) production model. While occasionally ships are produced in series, the end product is typically distinct. This uniqueness can be attributed to the frequent design alterations that occur throughout the lengthy production process. Moreover, even within the same series, customers often present varying specifications and requirements, which contribute to the uniqueness of each ship(Dugnas and Oterhals 2008).

Fixed position layout

In contrast to the conventional material flow observed in mass production, where materials move through different workstations, shipbuilding typically employs a fixed position approach. Vessels are not rooted in the same place, as many other fixed position production, but the vessels are too big to be moved around. This leads to staff, materials, tools, and equipment needing to be moved around in different areas, and the workstations are moving through the product (Dugnas and Oterhals 2008).

Temporary organizations

Consequently, shipbuilding, being an ETO production, is inherently a project-driven industry. There is a requirement to manage each specific project, and as such, temporary organizational structures are often established. Shipyards typically employ a workforce comprising individuals on both long-term and short-term contracts to effectively cater to fluctuating demands (Liker and Lamb 2000).

Consistent production facilities

Vessel construction takes place within the same facilities at the shipyard. This opens the possibility of having a good production infrastructure. After the hull is assembled its dock outfitting before launching the vessel and continuing with quay outfitting (Semini, Brett, Hagen et al. 2018).

Generic shipbuilding production processes

The generic shipbuilding production processes got broken down by Andritsos and Perez-Prat (2000). They came up with twelve main processes and two supportive processes performed for the main processes.

Generic shipbuilding production processes	
Main	1. Raw material reception and preparation 2. Marking, cutting and conditioning of steel plates and profiles 3. Fabrication of 2D blockss: welding of flat and shaped sub-assemblies (panels and sub-blocks) 4. Fabrication of 3D blocks in workshop 5. Pre-erection: assembly of 3D blocks and subassemblies into erection units 6. Prefabrication of pipes, supports, modules 7. Pre-outfitting 8. Blasting and painting/coating 9. Erection and outfitting in the dry-dock or slipway 10. Outfitting in dock (incl. piping, wiring, machinery etc.) 11. Finishing and outfitting onboard the floating vessel 12. Commissioning and sea trials
Supporting	13. Transport and handling 14. Dimensional control and inspection

Table 3: Generic shipbuilding production processes (Andritsos and Perez-Prat 2000).

3.4.1 Shipbuilding performance bench marking

Pires Jr et al. (2009) state that the basic criteria for evaluating the performance of a shipyard from the competitiveness point of view are: production cost, building time, and quality.

Production cost
The predominant factor influencing the price competitiveness of shipyards is the production cost, with labor cost often showcasing the most significant disparities between different shipyards. While materials and equipment are typically sourced from global markets—offering relatively uniform conditions for all builders—the primary variances in direct costs for shipyards constructing similar vessels can typically be traced back to labor costs. The cost of labor is dictated by the number of man-hours utilized and the cost per labor unit. Given that the labor unit cost, or wage cost, is typically defined by the country or region, the performance of a shipyard in terms of labor cost is primarily determined by its labor productivity (Pires Jr et al. 2009.

Shipbuilding is an industry that requires substantial financial resources. In a typical European-style shipbuilding endeavor, approximately 70% of the project's value-added might originate from sources outside the shipyard. In certain instances, this external contribution could even rise to as high as 85% (Hagen and Erikstad 2014).

Building time
The timeframe between a shipbuilding contract's signing and the ship's delivery is a key measure of competitiveness. It can directly affect if new contracts are won or not. This is primarily influenced by two factors: the volume of orders affecting berth occupancy and the time taken to obtain crucial inputs, with the main engine often being a deciding factor in current markets. The portion of the delivery timeline is most reflective of a shipyard's efficiency is the building time (Pires Jr et al. 2009). In countries with high labor costs their offshoring strategy and what yards are chosen for steelwork will also affect the lead time (Semini, Brett, Hagen et al. 2018).

Quality
Competitiveness in shipbuilding hinges on three core elements: cost, delivery time, and quality. However, it's important to understand that quality extends beyond the product itself to a more comprehensive consideration of market demands. This broader concept of quality encompasses is based on the definition of Pires Jr et al. (2009):

1. The ship's quality is evident in both direct and indirect maintenance expenses, lifespan, and resale value.
2. The shipyard's adaptability and technical prowess to cater to specific owner requirements.
3. The provision and effectiveness of after-sales support and warranty services.
4. The expectation for fewer supervisory staff during the build period, signaling greater reliability and reduced costs for the owner.

shipyard's competitive performance can be gauged using indices tied to these three fundamental criteria. Moreover, technical efficiency can be evaluated by comparing these indices to the volume of resources used or available. These resources encompass not only direct production inputs but also factors that broadly impact production, including those related to the industrial climate of a specific country or region and the unique production patterns of each shipyard (Pires Jr et al. 2009).

4 Norwegian shipbuilding industry

This chapter presents the current state of operation management at some of the larger Norwegian shipyards. The content presented is based on a literature study and interviews.

4.1 Characteristics

Norwegian yards are typically compact facilities with short distances for workers, material, information, and equipment. They have a long history of producing advanced ships and are effectively organized (Semini, Brett, Hagen et al. 2018). Norwegian yards tend to focus on building customized and complex ships and this leads to their product often are build as separate projects (Hagen and Erikstad 2014).Experts highlight that Norwegian shipyards are chosen for their proven adeptness in handling complexity and innovation, with a focus on integrating sophisticated systems in vessels, coupled with a reputation for reliability and quality delivery.

In every yard, a limited set of resources is shared among various projects, making it challenging to optimize resource utilization and sustain a seamless workflow. These resources comprise personnel (such as welders, plumbers, electricians, and carpenters), equipment (cranes, workstations, and material handling devices), and infrastructure (docks, halls, and storage spaces). Flow refers to the efficient, well-coordinated movement of materials, resources, and information (Resyard 2023.

Serveral of the larger Norwegian yards operate as manufacturers within project-based organizations, emphasizing the successful execution of individual projects and excelling in the production and delivery of highly customized, technologically advanced, and innovative products. However, they have not yet fully capitalized on resource efficiency and smooth flow. Workflows are not necessarily streamlined, workers can in some cases engage in excessive walking, and resources are not necessarily utilized optimally. To attain resource efficiency and seamless flow, effective resource planning is crucial. This encompasses long-, medium-, and short-term scheduling and rescheduling of resources, resource investments, spatial organization of resources (layout), and resource balancing (workload balancing) (Resyard 2023.

Norwegian shipyards are primarily set up as engineer-to-order (ETO) companies (Hagen and Erikstad 2014). Being an ETO company means that all the processes starting from the design stage, are customer-driven(Olhager 2003). Some Norwegian shipyards have designs series customers can choose from, but due to slightly various specifications and requirements, the final vessels are unique (Dugnas and Oterhals 2008). A characteristic found through expert interviews is the constantly changing project scope and details. This separates shipbuilding from several other ETO industries, as customers can change the requirements throughout the project period. This contributes to the complex ETO environment of the Norwegian shipbuilding industry.

Since 2015, the oil price drop has affected the industry. This has led to a decrease in demand for offshore oil and gas vessels. The oil and gas industry stood for around 80-90% of Norwegian shipyard activity(Mellbye et al. 2018). This change has resulted in substantial economic losses and caused them to explore new markets, such as passenger and cruise vessels.

As a result of increasing global competition from countries with lower labor costs, Norwegian yards shifted their focus to producing hulls outside of Norway (Semini, Brett, Hagen et al. 2018). They now focus on outfitting the ship, such as installing machinery and pipes; cabling and electrical systems; ventilation, air conditioning, and heating. An important part is installing and commissioning leading equipment, such as propulsion, engines, and other complex systems. The extent of work executed in countries with lower labor costs varies, but due to readily available expertise and streamlined quality control, a significant portion of steel processing and welding is undertaken in Eastern Europe.

Semini, Haartveit et al. (2014) put the level of offshoring into four strategies; Norwegian production (I), Norwegian block outfitting (II), Norwegian dock outfitting (III), and Norwegian quay outfitting (IV). There are currently no known Norwegian shipyards performing strategy (I), but yards are performing one or more of the other strategies. The choice of Strategy affects what works and how much work is performed in the yards. Norwegian yard's main strategies are (III) and (IV), with a handful also using strategy (II).

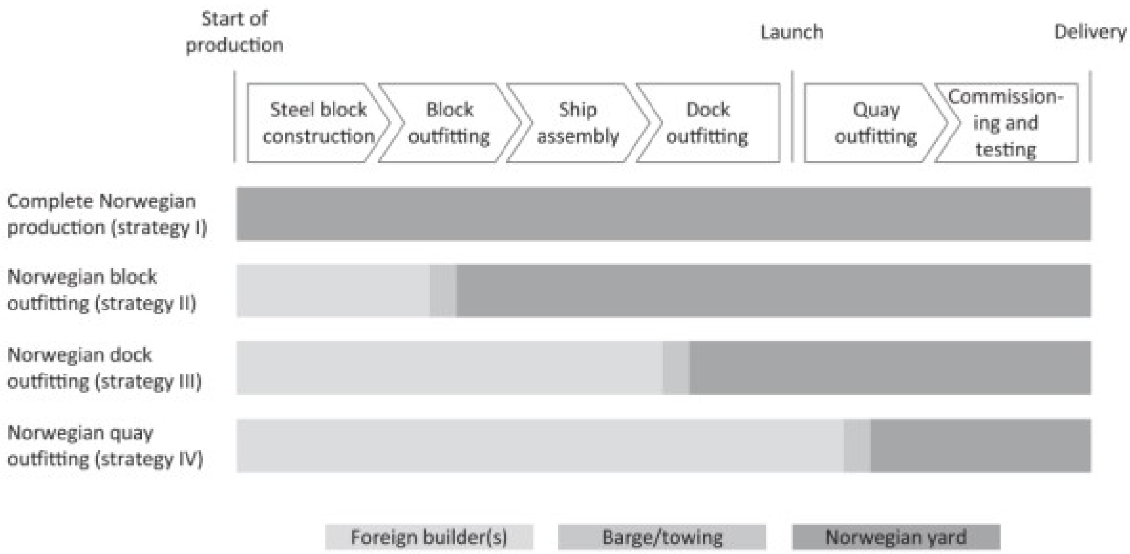


Figure 13: Four Norwegian ship production strategies, the difference being how much is performed at a foreign builder and a Norwegian yard (Semini, Brett, Hagen et al. 2018).

The main reason for the strategy choice is cost-related, as it's considerably cheaper to do steel work in eastern Europe. Another benefit of offshoring steel work is that the Norwegian yard has more time to complete parts of the drawings not needed for the steel work. However, it brings some disadvantages as well. Semini, Brett, Hagen et al. 2018 point to the main disadvantages of quality, delivery time, flexibility, and more stakeholders involved. The reasoning is as follows:

- *Quality:* As the Norwegian yard takes over the project at a later stage, it has less influence on the quality of the final product. This can be particularly problematic in the outfitting phase, as offshore builders may have less experience, and defects may not be detected until the project arrives in Norway, making quality control more challenging.
- *Delivery time:* The risk of delays caused by external factors increases with increased offshor-

ing. Additionally, the ability to recover lost time through local efforts (which may be costly) is diminished. Furthermore, the Norwegian yard needs more direct control over the progress at the foreign yard, and processes such as outfitting in closed ship structures tend to take longer in strategies III and IV.

- *Flexibility:* In shipbuilding, product flexibility refers to the ability to implement custom solutions and adapt to changes in design or specification. Norwegian shipyards have traditionally been known for their high product flexibility, and outsourcing production to foreign yards may compromise this capability.

- *More stakeholders involved:* Foreign shipyards may have less experience . This can be contrasted with Norwegian shipyards, which often have a higher level of experience and expertise in OSV production. Differences in the form of culture and language also need to be considered.

The automation degree in vessel outfitting is generally low, involving extensive manual labor. Strandhagen (2022) emphasizes the considerable progress needed to achieve fully digitalized yard logistics. Nonetheless, certain scenarios favor manual labor over emerging technologies. Groover (2008) mentions six situations where manual labor is preferred over automation:

- *Task is technologically too difficult to automate*: Reasons for the difficulty include (1) problems with physical access to the work location, (2) adjustments required in the task, (3) manual dexterity requirements, and (4) demands on hand-eye coordination.

- *Short product life cycle*If a product must be designed and introduced in a short period of time. In these cases, manual labor allows for a much sooner product launch than automated work.

- *Customized product*: Humans are more flexible than any automated robot and, therefore better if unique features are required.

- *Ups and Downs in demand*: An automated manufacturing system has a fixed cost associated with the investment. By downs in demand, the fixed cost is spread over fewer units of lead time.

- *Need to reduce the risk of product failure*: The ultimate success of introducing a new product is uncertain. Manual work reduces the risk of losing significant investments in automation if the product fails to meet the estimated market life.

- *Lack of capital*: If a company lacks the capital to invest in automated equipment, manual work is forced upon them.

4.2 Administrative processes in Norwegian shipyard

The current status of relevant administrative processes in a Norwegian shipyard will be presented, gathered from the literature study, expert interviews and information given by representatives from the case study yard.

4.2.1 Resource management

Shipyards are intricate and dynamic environments where a variety of resources are allotted for different operational activities, including labor (people), spaces (docks), machines, cranes, and other equipment. Each of these resources is essential to the processes involved in building, repairing, and maintaining ships, and the effectiveness with which they are used defines the shipyard's overall productivity and financial success (Hagen and Erikstad 2014).

Currently, shipyards manage resource allocation using a combination of manual labor, industry knowledge, and information systems. For instance, tracking continuous resource utilization is frequently done using an Enterprise Resource Planning (ERP) system. The shipyard management can track usage, schedule work, and make educated decisions thanks to the ERP system's role as a comprehensive database that saves and processes information about resources in real-time (Chen 2001).

However, experts report that when a new project is considered, detailed calculations are required to determine aspects like lead time, resource requirements and cost estimation. Due to the inherent variability and uncertainty in shipyard operations, these calculations—which are typically based on historical data and expert opinion—can be inaccurate.

Representatives from the case study shipuards says that the ERP system aids in determining current usage and future plans in terms of capacity checks. It does not, however, offer the capability to model situations in order to foresee probable obstacles or bottlenecks. This flaw is crucial, especially when making plans to change project schedules.

The effects of a project's need for acceleration or delay are normally calculated manually. Numerous variables are involved in this procedure, which has a big impact on the project's budget and schedule. The effects of schedule alterations are less foreseeable when there is no modeling capabilities, which could result in unforeseen expenses, delays, or resource shortages (cite: Kim2013simulation).

Although the usage of information systems like ERP helps the present shipyard resource allocation strategy, it still mainly relies on manual calculations and lacks simulation capabilities. This situation offers a chance for advancements that could further increase the effectiveness and efficiency of resource allocation in shipyards.

The approach towards labor as a resource shows some unique characteristics. These shipyards frequently use a high proportion of hired labor, which provides them with increased flexibility in managing their workforce (Hagen and Erikstad 2014). This strategy is particularly beneficial considering the cyclical nature of shipbuilding projects.

Throughout a shipbuilding project's lifecycle, different disciplines are mobilized at different stages. For instance, steelwork tends to peak before piping, which in turn precedes the peak of electrical work and outfitting. Consequently, the load curve for steel and piping workers is typically lagged within a project. Representatives informs that the current status of resource distribution strategy and estimation on what period the different departments are needed are mostly done manually with little help from information systems. These calculations and decisions are highly relied on experience.

By leveraging external, hired labor, shipyards can better balance their workload against the resources they are paying for. This approach allows them to avoid the potential inefficiencies and costs associated with having their workforce remain idle waiting for the next project. Therefore, in this context, labor hours are not perceived as a restricted resource but rather as an adjustable factor that can be scaled up or down based on the project's demands

Cost estimation
Accurate cost estimation allows shipbuilders to bid for contracts at a competitive price and to plan and manage the resources needed for a project (Hagen and Erikstad 2014). It also helps shipbuilders to identify potential cost savings and to avoid cost overruns, which can have serious financial consequences. In addition, accurate cost estimation is essential for ensuring that a ship is built within the specified budget and timeline and meets the required technical specifications.

The production cost is the main determinant of shipyard price competitiveness. The cost component that tends to present larger differences between shipyards is labor cost (Pires Jr et al. 2009). For this reason, the prediction of labor hours needed is essential for Norwegian yards to make an accurate cost estimation. Labor costs, inclusive of both internal and subcontracted resources, account for approximately one-third of the total costs. Furthermore, the three most substantial main system groups constitute 80% of the total costs, inclusive of services like installation incorporated within purchase costs. It is noteworthy that efficiency enhancements in the production process may yield reductions in procurement expenditures (Hagen and Erikstad 2014).

Lead time estimation
Evidence suggests that shorter delivery times may enhance a company's chances of securing contracts (Semini, Brett, J. O. Strandhagen et al. 2022). This could be due to the advantages customers gain from quicker deliveries, like more business opportunities and earlier cash inflows. In ETO manufacturing, swift responses often mean lower costs and greater efficiency. Also, experts highlight that reliable shipyards are sought after to avoid common vessel delivery delays. In case of a delay, the shipyard usually faces a customer fine or expensive overtime to meet the agreed delivery date. Thus, accurately predicting lead time can result in benefits such as improved customer satisfaction and cost savings.

The case study yard informed that lead time calculations are done with historic data from previous vessels and factors that describes the new potential project as size and complexity

Compensated Gross Tons (CGT)
The Compensated Gross Tonnage (GT) refers to the calculated volume, quantified in cubic meters, of a ship's interior hull and deckhouse, which is then multiplied by a particular factor (Lamb and Hellesoy 2002.

CGT (Compensated Gross Tonnage) is a measure used in the shipbuilding industry. It accounts for both the size and complexity of the ship and aims to take into account the difficulty, or the required workload, of building ships of different types and sizes (Pires Jr et al. 2009). CGT is an essential metric for comparing the productivity and output of shipyards, as it provides a standardized measure of the workload associated with building various types of ships.

The CGT calculation considers the ship's gross tonnage (GT), which is an indicator of the internal volume of the vessel, and a coefficient related to the ship type. Different ship types have different coefficients, reflecting the varying levels of complexity and labor required for their construction (Lamb and Hellesoy 2002). For example, building a tanker would require a different level of effort than building a container ship or a passenger vessel. The CGT is calculated as follows:

$$CGT = GTxCoefficient$$

4.3 Production processes at some of the larger Norwegian shipyards

With the different applied strategies by Norwegian yards, there are also differences in processes performed. The processes selected in this section try to give a picture of all the processes relevant to yards with strategy (III). The scope is Strategy (III) since that is the most used in the case study yard.

Using the production processes shown in table 4 as a foundation. Processes 2,3,4,5 can be eliminated because yards with strategy (III) have most of their steel work performed in eastern Europe (Semini, Brett, Hagen et al. 2018). As most steelwork is performed elsewhere it is a limited amount of raw material reception and preparation on the yards. However, there are some related to especially pipes. Process 12 in table 2 is performed before a vessel is delivered, but not really a production process and is therefore also eliminated. Figure 14 is showing the production processes that are performed in a shipyard following strategy (III)

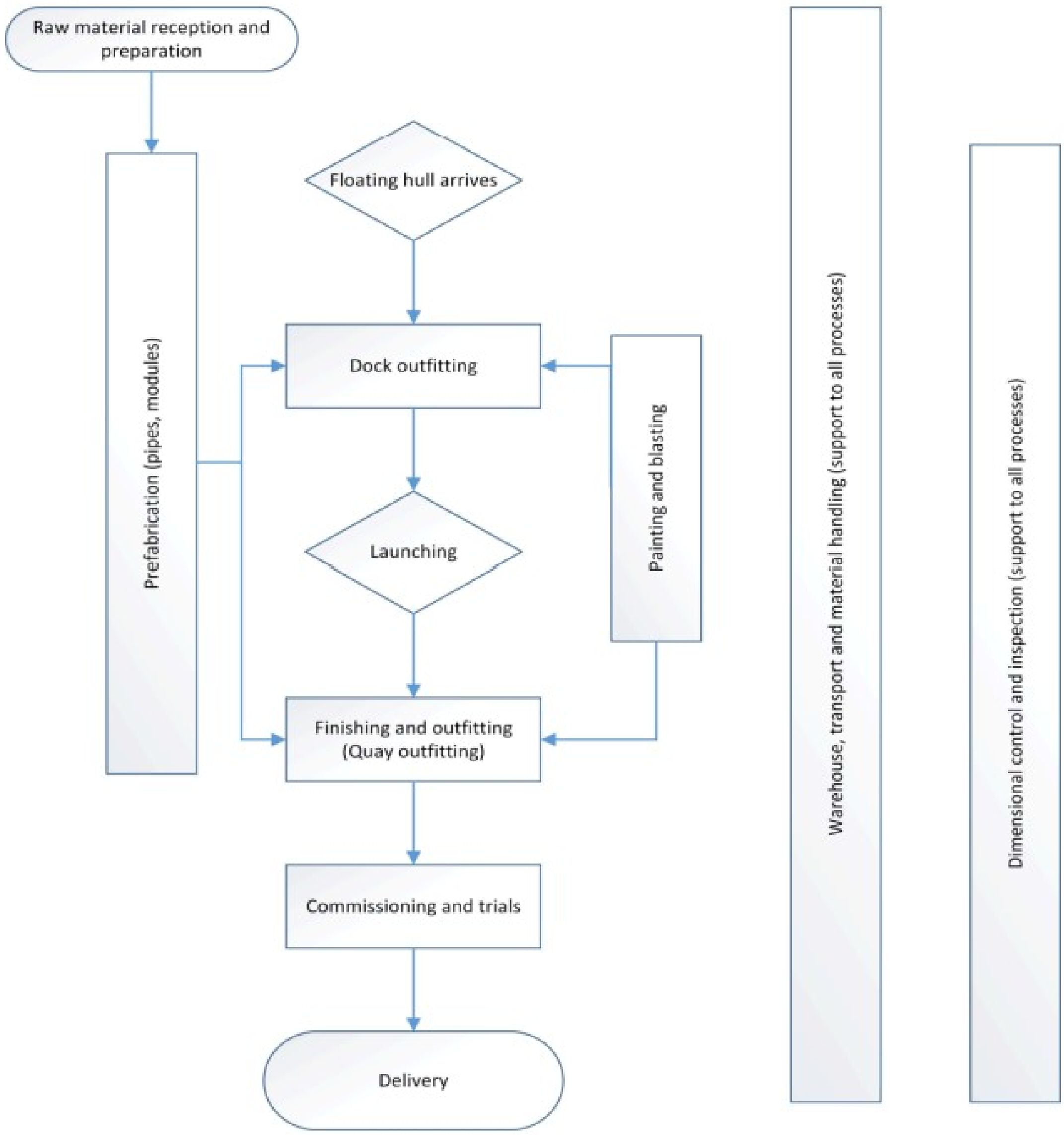


Figure 14: The production processes that are performed in a shipyard following strategy III (Hjartholm 2019).

4.4 Case company

The case company in this study has chosen to remain anonymous. Nonetheless, it is representative of a typical small, compact Norwegian shipyard, well-integrated into the local community. The shipyard is a significant economic player in the region, with each worker at the yard generating additional employment opportunities. Specifically, it has been estimated that every worker at norwegian yards indirectly creates an additional 5.4 jobs in the private or public sector (Oterhals et al. 2011). This illustrates the shipyard's central role and importance in sustaining and driving the local economy.

In the operational dynamics of the case company, a significant proportion of the workforce comprises subcontractors. The company's management strategy does not typically view manpower as a restricted resource, indicating a flexible approach to staffing and resource allocation.

The representative from the shipyard explains that they have available data, mostly located in administrative operations. They are very optimistic about researching the potential of the digital twin concept in their operations as they are always on the look of evolving. Implementation of virtual reality has been tested before, which shows a willingness to be an industry leader in the form of a technological revolution. As it is the only yard talked to during this research, it influences the research by concentrating on their current situation and not the current status of Norwegian shipyards in general at some points in the research. Their current status aligns with the information given in 4, but the different operations might be executed differently than described for Norwegian shipyards in general as information on their specifically execution was not possible to obtain at this point.

4.4.1 Design portfolio

Historically, the shipyard constructed a variety of vessel types; however, prior to the 2014 oil crisis, the primary focus was on vessels for the offshore oil and gas industry. Following the decline in oil prices, the demand for such vessels diminished, necessitating adaptation. Consequently, the design portfolio has become more diverse in recent years compared to its previous state. As of 2019, the design portfolio comprises the following:

- Offshore wind
- Cruise
- Offshore Oil and gas
- Fishing
- RoPax

As a consequence of market shifts, the shipyard now produces vessel types with which it has limited prior experience (Hjartholm 2019). While the company previously possessed extensive expertise in constructing technologically advanced offshore support vessels (OSVs), the transition to building passenger vessels and other segments has introduced differing requirements compared to OSVs. This has led to new challenges in areas where the company could previously rely on the experience and expertise of its personnel, which is no longer possible to the same extent.

In conjunction with other Norwegian shipbuilders, has historically specialized in creating highly customized vessels tailored to each client's specific needs. Under this approach, the engineering and design of a vessel would commence only after the client's involvement in the process. Presently, they have adopted a multi-strategy approach, offering both bespoke designs and a selection of standardized designs. These standardized designs feature established concepts, with a significant portion of the design and engineering work completed prior to identifying a specific customer (Semini, Haartveit et al. 2014). The vessels are categorized into first of a kind, Similar as previous vessel, Equal to previous vessel.

They focus on outfitting the ship, such as installing pipes and machinery; cabling and electrical systems; heating, ventilation, and air conditioning; and accommodation and hotel functions. A significant part of their steel work is done in countries with lower labor costs, strategy (III). With the high labor cost in Norway, the production of ships must be as cost-effective as possible.

5 Strategic areas of deployment of digital twins in shipyards

This section will explore possible strategic areas of deployment of digital twins in shipyards. The objective of this examination is to ascertain the current state and identify the optimal points of inception for integrating the digital twin concept within Norwegian shipyards.

In recent years, the concept of a Digital Twin (DT) has emerged as a transformative technology capable of revolutionizing industrial processes through its ability to mirror physical systems in a digital environment. This powerful tool, which integrates real-time data with predictive modeling and artificial intelligence, offers potential solutions for a myriad of industrial challenges, from optimizing production lines to predicting equipment failures before they occur (Kritzinger et al. 2018). However, the practical application and implications of this technology within the shipbuilding industry, specifically within Norwegian shipyards, have yet to be examined within the existing literature.

Wang et al. 2022 did a literature review of digital twin applications for process management in the ship industry in 2022. Nine applications was identified, where only three of them were directly linked to shipyard operation and two of them were linked to hazard and risk management.

Given the, in my knowledge, limited literature on Digital Twin applications in shipyards, it's essential to extract insights from other industries where DT has been successfully implemented. Industries like manufacturing and logistics, which share operational similarities with shipyards, demonstrate how DT can enhance efficiency, cost-effectiveness, and predictive capabilities. Therefore, despite the current lack of shipyard-specific research, the successful deployment of DT in other sectors provides a compelling case for its potential applicability and value in the shipyard industry.

Industry professionals and representatives from shipyards frequently highlight the high initial costs of adopting new technologies as a major obstacle. This perspective is also shared by representatives from the case company, who emphasized the strict evaluation standards for potential new investments. They stressed that any proposed investment must demonstrate a significant potential for positive effects on the yard's key performance indicators before it can be considered for implementation. Norwegian shipyards can effectively address the cost challenges associated with digital twin technology by focusing on specific application areas.

With a incremental approach where implementation of DT technology in areas with the highest anticipated benefits, shipyards can initiate the digital transformation journey. The gradual integration and subsequent expansion of these applications into a comprehensive system allow for a phased approach. This method not only enables shipyards to maximize the advantages of digitization but also minimizes potential cost and complexity challenges associated with sudden, large-scale implementation. In this way, shipyards can effectively manage their digital transition, ensuring both operational enhancements and cost efficiency.

In semi-structured interviews conducted with representatives from the digital twin department of a renowned consulting firm, they confirmed that this incremental approach is a well-used methodology. The incremental approach of starting with specific applications and gradually expanding is generally regarded as practical and effective. This allows organizations to manage the initial costs, mitigate risks, and learn from smaller-scale deployments before fully committing to more complex or widespread applications.

Typically, starting with areas where the most immediate and significant benefits can be observed is a common strategy. From there, organizations can build upon their initial successes and gradually expand the use of digital twins to other areas based on their specific needs, capabilities, and strategic objectives.

5.1 Possible digital twin applications for shipyards

This sub-chapter will go more into the depth of some of the potential areas where applications in the digital twin concept can be introduced. It's important to keep in mind that most suggested application are not proven in the shipbuilding industry, and even with many potential benefits, there are also known challenges and probably manage unidentified challenges to overcome.

Yard layout and logistics optimization
Yard layout optimization refers to the process of designing or reorganizing the layout of a shipyard, industrial facility, or any other operational space in a way that maximizes its efficiency, productivity, and overall effectiveness. The goal is to create a streamlined and well-organized layout that minimizes operational costs and resources while maximizing throughput (Stevenson et al. 2014). Digital twins enable shipyard managers to analyze and optimize the layout of the yard, identifying the most efficient arrangement of equipment, resources, and storage areas to minimize wasted space and time.

Yards in Norway are project-oriented manufacturers, excelling in executing individual projects effectively and producing highly customized, technologically advanced, and innovative products. However, they have yet to fully optimize resource efficiency and workflow. Current processes are not streamlined, with workers often engaging in excessive movement, and resources not being utilized to their full potential. To improve resource efficiency and workflow, strategic resource planning becomes crucial. This includes physical arrangement of the layout (Resyard 2023).

Using real-time data with digital technologyies such as machine learning algorithms to optimize the routing of materials within a shipyard can improve efficiency by reducing the time and resources required for material movement (Alhaidari et al. 2021). By training the algorithm on data such as material location, ship component location, and shipyard layout, the most efficient routes for material movement can be identified. The algorithm considers distance, vehicle and worker availability, and potential obstacles to avoid delays and minimize resource usage. The potential benefits of using machine intelligence in routing optimization include improved network performance and the ability to satisfy better user demands, such as scalability, mobility, reliability, and quality of service (Dai et al. 2021). Varghese and Tamma 2018 did a study on the layout of outfitting shipyards. This study shows the complexity and vast amount of variables that need to be considered when optimizing the routing of materials within a shipyard.

A digital twin of a shipyard can help in improving various key aspects of yard layout optimization, such as:

1. Material flow: A digital twin can simulate the flow of materials throughout the shipyard, identifying bottlenecks and areas of congestion. By analyzing this data, stakeholders can make informed decisions to streamline material flow and reduce handling times (Rosen et al. 2015).
2. Space utilization: The digital twin can help visualize and analyze the spatial utilization of the shipyard, enabling stakeholders to identify underutilized or overcrowded areas. Adjustments can be made to the layout, ensuring optimal use of available space and efficient operations.
3. Equipment placement: Using a digital twin, shipyard managers can simulate and evaluate different equipment placement scenarios. This helps in identifying the most efficient positioning of machinery and other assets, minimizing movement and improving overall process efficiency.
4. Workforce efficiency: A digital twin can help identify areas where workers' travel time and effort can be minimized by simulating worker movements and tasks. Adjustments to the layout can be made to ensure a more efficient working environment, leading to increased productivity (Kunkera et al. 2022).
5. Safety considerations: By incorporating safety regulations and guidelines into the digital twin, shipyard managers can identify potential hazards and ensure compliance with safety standards. This helps minimize the risk of accidents and maintain a safe working environment (Kunkera et al. 2022).
6. Flexibility and adaptability: The digital twin can simulate future demands and trends, helping stakeholders proactively plan and adapt the yard layout to accommodate anticipated changes. This ensures the shipyard remains flexible and adaptable to evolving requirements.

Resource and capacity management

In the context of shipbuilding, the concept of executing tasks correctly is predominantly tied to adequate preparation (Hagen and Erikstad 2014). The production of large, advanced ships in contemporary shipbuilding hinges on a professional support infrastructure. This infrastructure ensures that the production sector of the organization is well-informed about what needs to be built, when it should be built, and how it should be constructed. Simultaneously, it ensures the availability of all necessary resources to effectively carry out the tasks.

A comprehensive digital twin for resource and capacity management in a shipyard would represent a highly sophisticated, interconnected system that meticulously mirrors every physical facet of the shipyard, incorporating real-time data on resource allocation, labor hours, equipment availability, and capacity constraints. Such a digital twin would coalesce seamlessly with planning and operational tools, yielding comprehensive visibility over the entire shipyard operation, thereby facilitating more informed decision-making processes. This could also make the resource management take advantage of the simulation aspect of the digital twin that Rosen et al. (2015) talks about. With the simulation on real-time data stakeholders could simulate various scenarios, such as changes in resource allocation, workflow adjustments, or facility expansions. These simulations can help identify the optimal balance between resource usage, capacity, and cost, leading to more efficient and effective shipyard operations.

In these simulations, there would be applications such as lead time prediction, cost estimation, prediction of resource usage, demand forecasting, and process tracking. These are all applications that are been proven in similar industries:

- *Lead time:*Lingitz et al. 2018 uses different machine learning algorithms to show the possibilities of doing lead time prediction and analysis on a semiconductor manufacturer. Semini, Brett, J. O. Strandhagen et al. (2022) did a regression analysis on a dataset of vessels built in Norway and found a correlation between several parameters, such as the ship's size, to affect a vessel's lead time.
- *Cost estimation:* Kaluzny et al. (2011) did a study using data mining algorithms to predict the cost of a ship based on the ship's characteristics. It was concluded that the algorithms can be effectively used for shipbuilding cost estimation and could improve the accuracy and efficiency of cost estimation in the shipbuilding industry
- *Process estimation:* Process tracking using machine learning has been used in the ETO environment for the last few years. Openspace has released a machine learning-powered progress-tracking solution. It uses computer vision to turn pictures into millions of valuable data points to help better track your project's progress (Hedmond 2020).
- *Demand forecast* Dou et al. 2021 showed an example of this when they could, with high accuracy, predict the demand in the manufacturing industry with the help of machine learning. The results show that prediction based on historical data alone is successful, but the accuracy of the prediction results is significantly lower than when considering multiple factors. This points to the implementation of real-time data from a digital twin would produce a sufficient forecast.

Digital twin technology serves as the basis for various of these applications aiming to perform predictions, estimations, and forecasts. By integrating this technology with real-time data within a digital twin, enhanced performance can be achieved. This finding is corroborated by the study conducted by Wang et al. (2022), where the authors investigated the operation within the engine room of a vessel and confirmed the higher performance afforded by the conjunction of real-time data and digital twin technology.

Upon the entry of new potential project features, this ideal system would automatically analyze the requirements, predicting the resources needed for each project phase, from conceptual design to delivery. This prediction would encompass everything from raw materials, and manpower, to machinery hours, considering concurrently running and future projects, accounting for their resource requirements and timelines to provide an accurate portrayal of resource availability and potential conflicts. In the case study in the next chapter, there is a description of predicting the manpower needed for new projects.

Digital twins can monitor and analyze resources as energy consumption, water usage, and waste generation, helping shipyards identify opportunities for reducing environmental impact and lowering operational costs. Kunkera et al. (2022) implemented the digital twin concept into a shipbuilding project and found that compared to a similar vessel, the energy consumption went down. The study showcased the significance of 3D modeling in a virtual environment through a comparison of the construction and outfitting processes of a passenger vessel and a cruise vessel. The use of a 3D model or Digital Twin allowed outfitting at the earlier stages of construction, improving project realization productivity by about 20% and enabling a 20% reduction in electrical energy consumption. This has also led to a decrease in CO_2 emissions by 1140 tons over two years.

Disaster preparedness and emergency response

Digital twins can simulate various disaster scenarios and help shipyards develop effective emergency response plans, ensuring the safety of workers and assets in the event of an incident. Wang et al. (2022) investigated two examples of the DT approach, revealing that the DT system can help decision-makers in devising advanced response plans for emergencies in the ship operation process. Kunkera et al. (2022) did a case study with the use of the digital twin concept in shipbuilding and compared it to a similar vessel without the digital twin technology. It was found that safety improved, with a reduction in injury frequency by about 43% and a five-fold reduction in injury severity due to advanced outfitting.

Predictive Maintenance and Crane and Equipment Utilization

Another promising application area for digital twins in shipyards is in the realm of predictive maintenance and utilization of various equipment. By collecting and analyzing data from sensors installed in the shipyard's equipment and infrastructure, a digital twin can monitor the performance of critical components and systems, predict potential failures, and suggest maintenance actions before a problem occurs. This can help shipyards minimize downtime, reduce maintenance costs, and increase the overall reliability and lifespan of their facilities.

When it comes to maintenance, the traditional reactive approach – performing repairs only after a failure has occurred – can lead to unplanned downtime, inefficiency, and excessive costs. Proactive maintenance strategies, such as preventive and predictive maintenance, have proven to be far more cost-effective and efficient (Lee et al. 2013). Aivaliotis et al. (2019) used a digital twin system successfully to estimate the Remaining Useful Life (RUL) of a defective gearbox in a robot, providing accurate results beneficial for predictive maintenance and production planning decisions. Süve et al. (2022) introduce a framework on how to implement such an application with only the use of the digital twin technologies IoT and ML. The framework provides a digital twin simulation of the production environment integrated with the real world and the ML models to evaluate the effect of different parameters, such as the throughput rate or the cost.

Modern shipyard operations heavily rely on cranes and other heavy equipment (Hagen and Erikstad 2014). These resources are a significant investment, and their effective utilization is a key determinant of operational efficiency. By mirroring the operational status of these assets in real-time, digital twins enable a more precise and dynamic understanding of equipment performance and usage patterns. Wang et al. (2022) performed a case study that proves that predictions in the engine room are better with real-time data. An improved Bayesian Neural Network (BNN) algorithm is used to optimize critical parameters in the ship operation process, enabling dynamic prediction of failure and risk. This, in turn, allows for data-driven decision-making aimed at optimizing equipment utilization and reducing potential downtime.

6 Case Study on resource management in the Shipyard

Several potential areas for the case study, some outlined in section 5, were proposed to the shipyard under investigation. Among the various options, the shipyard expressed a particular interest in a study focused on resource management, specifically forecasting labor hours. This choice was justified by two main considerations. Firstly, labor constitutes a critical operational aspect, accounting for approximately a third of total costs (Hagen and Erikstad 2014). Secondly, the requirements for implementing such a study are met at the shipyard. It is believed that data of sufficient quality is available, and the investment costs are minimized due to the technical expertise being provided by a student at this stage and only existing systems are used. A case study on resource allocation by the use of information systems at a Norwegian shipyard, focusing on the digital twin concept, is an opportunity to explore some of the benefits of digitalization in the shipbuilding industry.

The shipyard, a renowned shipbuilding company, has encountered challenges in resource allocation, particularly in terms of manual labor hours for potential projects. The current methodology requires manual estimation of the necessary labor hours, which are subsequently incorporated into their visualization system, Tableau. This system interfaces with their ERP platform, allowing for a comprehensive visualization of labor hours required for all ongoing projects. Despite its utility, the manual operation of labor estimation is subject to human error and inefficiencies. Representatives from the case company point out the calculation of labor hours needed for a new potential project as a crucial and complex operation. Given that labor hours constitute a substantial proportion of the overall shipbuilding cost, accurate estimation, and optimization are crucial to maintain economic viability within the shipyard.

In light of these issues, The yard seeks to develop an automated predictive model for labor hour estimation. The proposed system would employ a data-driven approach, utilizing project-specific factors to predict required labor hours automatically. Integration of this model into the Tableau visualization system could eliminate the need for manual estimation and reduce inaccuracies, thereby streamlining the allocation of resources. The enhanced system is anticipated to provide more accurate labor hour estimates, enabling management to make better-informed decisions and optimize resource allocation. Consequently, this approach has the potential to decrease project lead times and increase overall efficiency by identifying the factors affecting the amount of labor hours. Furthermore, the seamless integration of the predictive model with the existing ERP system and Tableau visualization tool ensures ease of use and uninterrupted data flow.

This proposed solution aligns with the digital twin concept, wherein a virtual replica of a physical asset or system is generated to optimize performance, maintenance, and resource allocation. By integrating the automated predictive model into the yard's existing infrastructure, the digital twin of their shipbuilding projects will be enriched with automatic labor hour estimates. This, in turn, enables real-time monitoring and assessment of resource allocation, allowing for proactive adjustments to optimize project management. The implementation of the digital twin concept within the yard's operations demonstrates the potential of leveraging cutting-edge technology to enhance decision-making processes and overall project efficiency in the shipbuilding industry.

6.1 Data collection

A field visit was conducted at the shipyard for the purpose of data collection, with the primary aim of gathering comprehensive historical data related to the construction of various vessels. This visit included an interview and meetings with representatives from the shipyard, who provided invaluable insights into the operations and processes undertaken during ship construction. These interactions offered a practical perspective, complementing the quantitative data acquired and helping to create a holistic understanding of the shipyard's operations.

The data utilized in this case study was generously provided by the shipyard and comprises of a historic dataset of all vessels built and/or outfitted at the yard. This data encompasses multiple factors associated with each vessel, aligning with those that are typically considered when a potential new project is initiated. The factors include but are not limited to the vessel's specifications, the design complexity, equipment installed, and the labor hours required for construction. By examining these variables, we can gain a deep understanding of the processes and efficiencies in the shipyard, ultimately informing the implementation of Digital Twin

The case study yard wants prediction on hours needed for the following departments:

Department	**Description**
Total hours	Total hours used on vessel
Scaffolding	Installation, maintenance, and removal of temporary support structures
Piping & machinery	Installation, and integration of piping systems and mechanical equipment
Outfitting yard	Total time used at the outfitting yard
Production outfitting	Total time used for outfitting
Engineering	Total time used by the engineering department on a vessel
Management	Time used by management department on a vessel

Table 4: Departments and descriptions

6.2 Analyzing of data

In the first step of the case study, the focus was on the analysis of data provided by the yard. The dataset received from the yard consists of 59 examples of vessels built, with each vessel connected to a variety of factors, such as CGT, cabins, series, compactness, engine power, and other attributes. In addition, the dataset includes information about the number of hours spent in each department during the vessel construction process. The dataset is highly dimensional, with numerous factors and large amounts of data for each vessel. The data will be analyzed to identify patterns, correlations, and relationships between the factors and the time spent in each department.

There isn't a strict definition for what constitutes a small dataset in machine learning, as it can depend on various factors like the problem domain, complexity, and the model being used. However, a rough guideline would categorize fewer than 10,000 samples as small (Goodfellow et al. 2016). By this definition, the dataset at hand can be labeled as small.

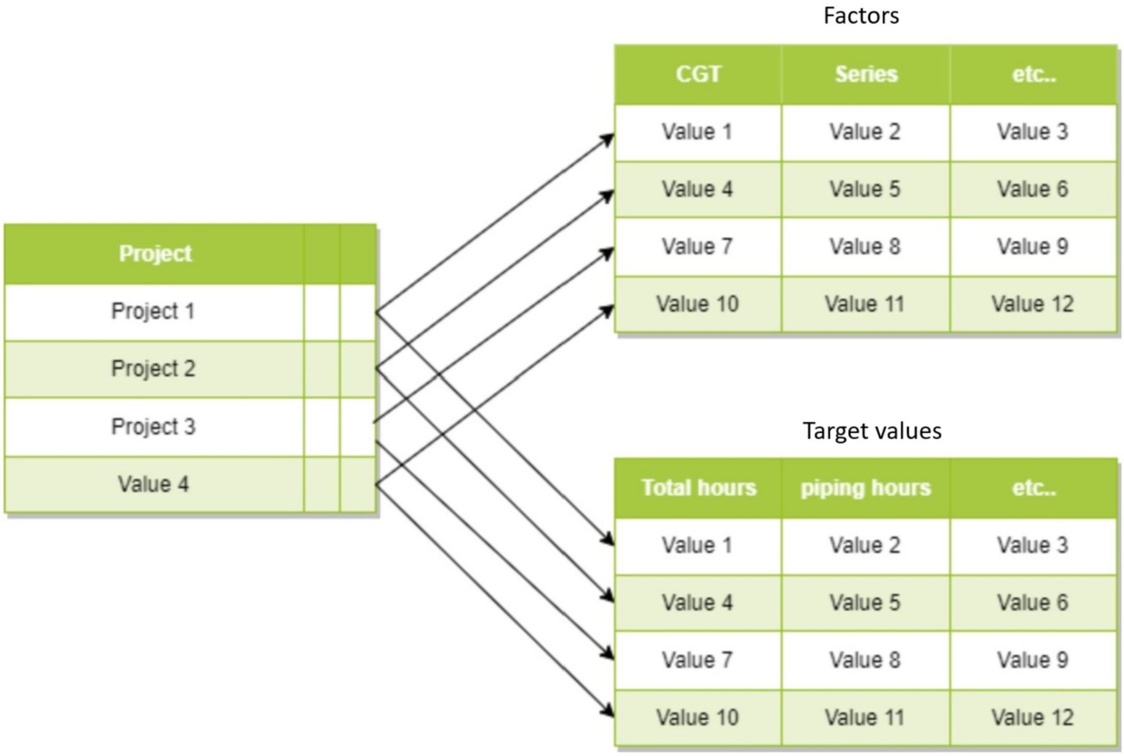


Figure 15: Data structure case study

Before presenting the analysis, the data was first inspected and cleaned to ensure that it was in a usable format. This involved checking for any missing or erroneous values, inconsistencies, or duplicate records that could potentially affect the results of the analysis. Outliers was also looked into, which could be genuine data points or the result of errors in data entry or measurement. No significant issues were found, and after carefully addressing a few insignificant issues, the analysis proceeded. Some outliner was found but kept in the dataset as outliers in a small dataset should not be deleted because they may carry valuable information that can significantly impact the overall analysis. Deleting them could lead to biased results and misinterpretation of the dataset, as the remaining data may not provide a comprehensive representation of the underlying patterns or trends.

To enable the use of machine learning models for analyzing the dataset received from the yard, the textual factors, such as "Series" which had values "First," "Similar," and "Equal," were transformed into numerical representations. This was necessary because machine learning algorithms require numerical data as input. Therefore, each unique value in the textual factor was assigned a corresponding numerical value, with "First" being assigned a value of 1, "Similar" a value of 2, and "Equal" a value of 3. This conversion process is known as factor encoding and is a common technique used in machine learning to convert categorical data into numerical data that can be used in models (Pedregosa et al. 2011). The numerical representation of factors is necessary for the analysis of data using machine learning algorithms, such as clustering or regression, that are designed to work with numerical data. The same was done to Building strategy, where the numerical value matches the strategies defined by (Semini, Haartveit et al. 2014) and presented in figure

13.

Next, a descriptive statistics on the dataset was performed, which included calculating measures of central tendency (mean, median, and mode), dispersion (range, variance, and standard deviation), and shape (skewness and kurtosis) for each parameter. These calculations provided valuable insights into the distribution and variation of the data, helping me understand the overall structure and characteristics of the dataset. The distribution of total hours divided on each department was also looked into to get an understanding of the labor allocation.

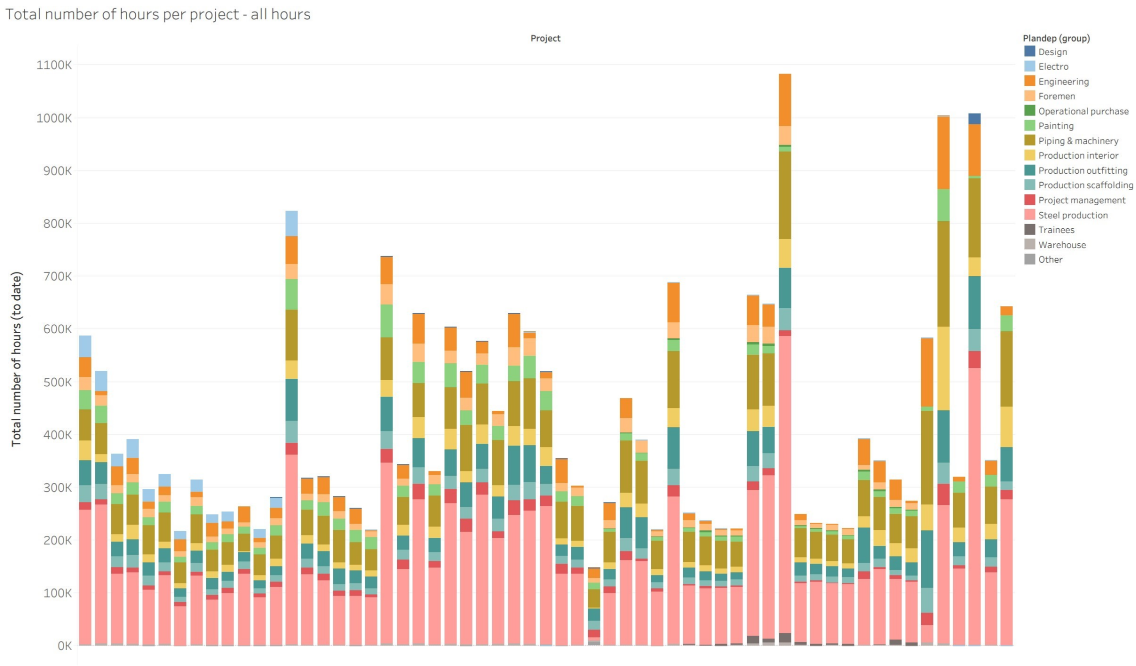


Figure 16: Visualization of hours used in each department on different projects

6.3 Selecting relevant factors

In the third step of the case study, the focus was directed toward factor selection. This pivotal process involves identifying the most pertinent parameters or variables that influence the number of hours required to construct a vessel. By selecting the appropriate factors, the accuracy and interpretability of the predictive model could be enhanced, which would ultimately assist the shipyard in making more informed decisions regarding resource allocation.

When receiving a new potential project, several vessel factors are attached to the project. From these factors the yard wants to predict the hours needed throughout the project. The available factors are the same for every new potential project. In addition, The historical dataset provided included the total hours used on each project, as well as a detailed breakdown of the hours utilized by each department involved in the shipbuilding process. This granular information enabled a comprehensive understanding of labor allocation and facilitated the development of a predictive model tailored to the yard's unique requirements and operational structure. A description of the factors is given in the table below.

Factor description

Factor	**Description**
Project	Newbuilding number at the yard
Series	First, similar or repeat
GT	Gross tonnage
CGT	Compensated Gross Tonnage
Cabins	Amount of cabins on vessel
Building stratagy	Building strategy used (1,2,3 or 4)
Vesselcategory	Vessel category (AHTS, PSV, etc)
Design	Design name
Compactness	CGT/GT
Engine KW total	Total installed power (excl .battery)
Engine Nos	Number of engines
Engine Type	DE/DM/Combi
No thrusters	The number of thrusters installed on the vessel
Settleweight	Steel weight
Accomodationarea	Accommodation area for crew and passengers
Outfitweight (w topside)	Outfit weight
DP index	Dynamic positioning class (0or1,2,3)
POB	People onboard
KmCable	The length of cable used for mooring the vessel
Density index	Outfitting weight / steel weight
Powerplant index	(Nos. engine+nos. Thrusters+DP ind+ Engine type)/Installed power
Powerplant index2	Nos. engine+nos. Thrusters+DP ind+ Engine type
Accommodation index	(Nos. Cabins x accommodation area)/POB
Market index	Nos. of vessels contracted worldwide per contract year
Density index2	Outfitting weight / steel weight
Segmentexperience index	Number of vessels of each type built at the yard since 2000
Experience index	Grouped Segmentexperience index

Table 5: Factor description

In the process of data cleaning and preparation for analysis, it is often necessary to handle missing data. When a significant portion of data is missing for a specific factor, the reliability and validity of any analysis involving that factor can be compromised (Wu 2013). In this project, the limit was set to drop any factor for which more than 50% of the data was missing. This threshold was chosen to ensure that the remaining data would be representative and provide a solid basis for further analysis. Hence, the variables 'KmCable' and 'TonnesPipe' were subsequently dropped from further investigative analysis.

The factor selection process began with an analysis of the correlations between the hours used by different departments on a vessel and the various factors associated with it. This involved calculating the correlation coefficients for each parameter and plotting them in a heatmap to visualize the strength and direction of the relationships. By selecting factors with strong correlations to the target variable, the model can better capture the underlying patterns and relationships in the data, leading to more precise labor hour predictions (Guyon and Elisseeff 2003). This approach can help narrow down the list of potential factors to include in the predictive model. Pearson correlation, also known as Pearson's r, is used in this study for correlation calculations. It provides a value between -1 and 1, where 1 represents a perfect positive linear relationship, -1 represents a perfect negative linear relationship, and 0 indicates no linear relationship (Field 2013).

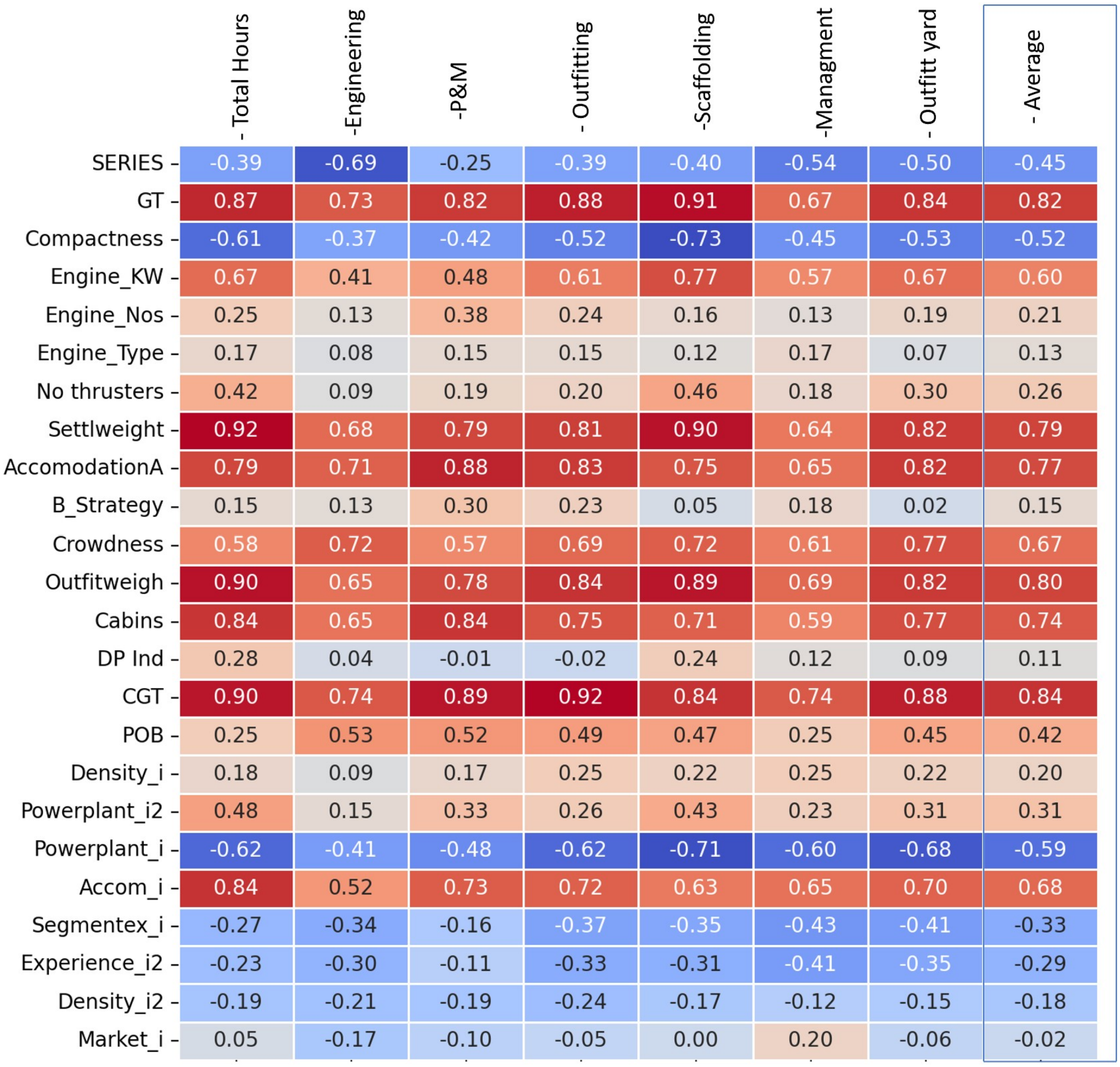


Figure 17: Pearson correlation between factors and hours used by different departments

This correlation table provides a quantitative measure of how each factor is related to the hours required in different departments. It is a valuable tool for understanding which factors might be the most important in planning and resource allocation for vessel construction. However, correlation does not imply causation, so these results should be interpreted with care and backed up with further analysis.

In this study, it was deemed essential to refine the factor selection based on the strength of their correlation with the dependent variable, hours used, across different departments. Factors demonstrating a correlation coefficient between -0.6 and 0.6 were perceived as bearing less substantive impact on the predictive capability of the models. As such, these factors were considered for elimination at this point of the analysis. The rationale for this decision was founded on the presumption that factors with low to moderate correlation are less likely to significantly influence the outcome variable, thereby potentially diluting the model's predictive power. Consequently, this pruning step is aimed at enhancing the model's efficiency and focusing on the factors that are most relevant to the prediction of hours used. The dataset is small with many factors and for that reason necessary to limit the factors before model development.

Furthermore, it is worth mentioning that the dataset includes categorical factors such as 'Series' and 'Building strategy'. Typically, the computation of correlation coefficients for these types of factors may not yield meaningful results due to their non-continuous, categorical nature. However, recent research conducted by Semini, Brett, J. O. Strandhagen et al. (2022) has demonstrated a meaningful correlation between these factors and the lead time of vessels. As such, in light of these findings, these factors are decided to retain in the subsequent analyses throughout the remainder of this study.

In the analysis of the given dataset, it is evident that weight-related factors such as 'GT', 'Settleweight', 'Outfitweight', and 'CGT' demonstrate a high degree of correlation with the total hours required across different departments. These factors exhibit strong correlations, with values exceeding 0.8 in most cases, indicating that they have a strong influence on labor hours. This suggests a significant relationship between the weight attributes of the vessel and the corresponding hours of work required, indicating that as the weight increase, so does the requisite labor.

Factors such as "Cabins", "Accommodationarea", "Crowdiness", "Accomodation-index", and "Engine KW" can also be linked to the weight of the ships. For instance, a bigger ship usually will have more cabins.

It's then also logical to think that they are highly correlated with each other.Introducing factors with high correlations into the same predictive model can lead to multicollinearity issues, which can negatively impact the model's performance and interoperability. Selecting only one or two of them can help prevent redundancy and simplify the model (Farrar and Glauber 1967). A correlation analysis with Pearson correlation is done with all factors linked to weight and not yet excluded. the figure below shows a heatmap that visualizes the pearson correlation between each of the factors related to weight. This was done through a Python script that can be found in Appendix B.

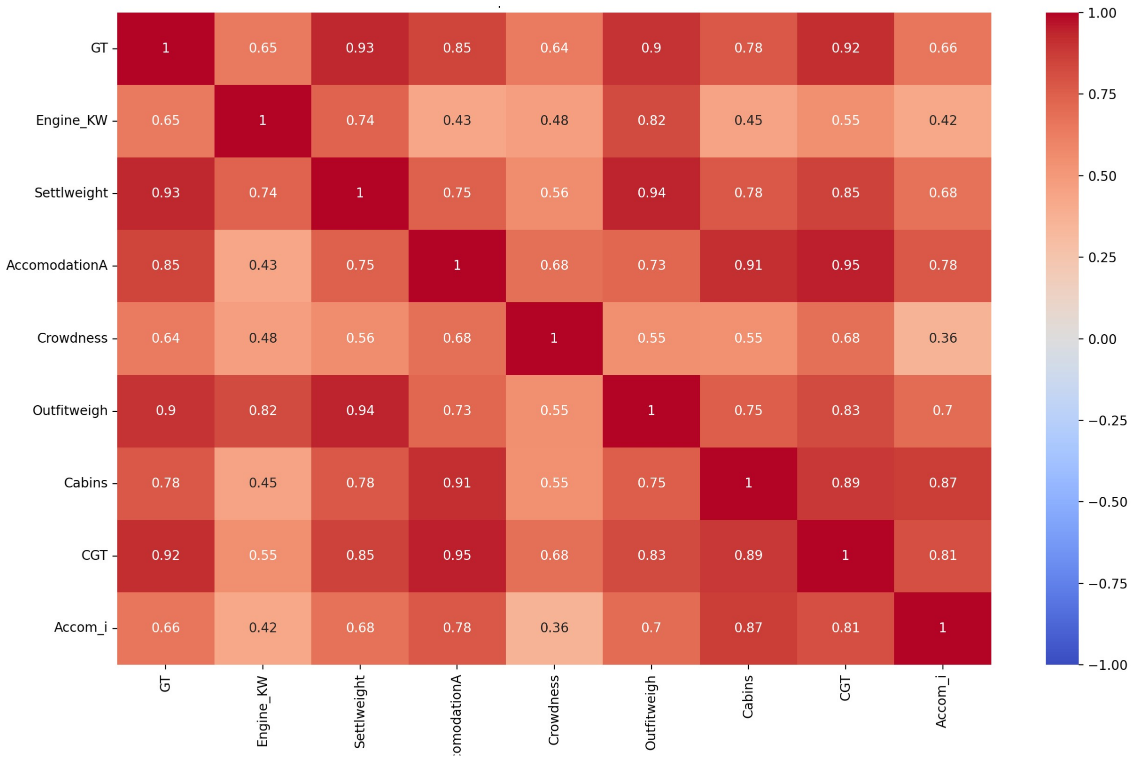


Figure 18: cross Correlation between all weight-related factors

The outcome of the correlation analysis revealed a pronounced interrelationship between all the examined factors. This strong correlation suggests that changes in one variable are likely to be accompanied by corresponding changes in the other variables. Thus, the robust linkages among all the factors suggest that they may interact significantly within the system under study. Among the factors assessed, "Engine KW" and "Cordiness" displayed the lowest correlation with the remaining variables. This observation could potentially be attributed to the variable requirements among different vessels. As a specific instance, identical engine power (measured in KW) might be implemented in two disparate vessels in terms of size. However, despite the common engine power, the smaller vessel may demand higher speed, and thereby lighter, thus contributing to the observed low correlation.

Following the correlation analysis, an analysis of factor importance was performed using a random forest model to further explore the significance of each factor. In this process, all weight-related factors were incorporated into the model. The random forest algorithm works by creating a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. During this training, the algorithm calculates a score that quantifies the importance of each factor in the prediction. This factor importance is calculated as the total reduction of the criterion brought by that factor, which is also known as the Gini importance (Strobl et al. 2007). Therefore, the higher the score, the more important the factor is in the decision-making process of the model, and all add up to 1. This analysis provides another layer of understanding beyond correlation, helping to identify not just relationships but also the relative influence of each factor on the predictive model. Such insights are crucial for model simplification and optimization, and they can guide factor selection for future modeling efforts.

factor	**Importance Score**
CGT	0.384
GT	0.222
Outfitweight (w topside)	0.143
Settleweight (w topside)	0.110
Accomodationarea	0.105
Cabins	0.034
Engine KW total	0.021

Table 6: Importance score of each factor in the Random Forest model

The table presented above delineates the factor Importance, based on Gini impurity, associated with each feature in the Random Forest model. These scores quantify the extent to which each feature contributes to the improvement of the model's predictions, by measuring the total reduction in the error brought about by that feature. Essentially, a higher score signifies a more substantial influence of the feature on the model's predictive capability. The CGT factor exhibits the highest importance score of 0.354, suggesting it has the most substantial influence on the model's decision-making process. An attempt to include all factors, except "Design" and "vesselcategory", into the random forest model was also done. In this model, "CGT", "GT", "Series" and "Outfitweight" emerged with the highest Gini scores, while the remaining variables presented scores so low that further differentiation was deemed nonsensical for this small dataset. However, it's essential to consider that the relatively small size of the dataset coupled with a large number of factors could introduce an element of randomness into the model's outcomes.

The decision to exclude these factors has been made based on several considerations: Their correlation with the target variables, their importance in the context of weight-related factors, and their overall importance according to the factor importance calculations. Notably, despite its strong performance, the "GT" factor has also been eliminated, primarily due to its inherent relationship with the "CGT" factor, the latter being a derivative of the former.

factor	Correlation with other factors	Negligible impact
Series		
GT	X	
CGT		
Cabins	X	
Building strategy		
Vesselcategory		
Design		
Compactness	X	
Engine KW total	X	
Engine Nos		X
No trusters		X
Settleweight	X	
Accomondationarea	X	
Outfitwegiht		
DP index		X
POB		X
Density Index	X	
Powerplant index		
Powerplant index2		X
Acoomodation index	X	
Market index		X
Density Indelx2		X
Segmentexperience		X
Experience index		X

It is important to keep these and all other excluded factors in mind when interpreting the results, especially the ones likely to be related to the factors included in the models.

In studying factors affecting the labor hours needed for shipbuilding, several key works provide insightful perspectives. Pires Jr et al. (2009) presented a novel methodology to identify efficient shipyards and calculate inefficiency scores for others using Data Envelopment Analysis (DEA) (Pires Jr et al. 2009. Their study prioritized partial productivity indices, specifically labor productivity measured in man-hours per CGT. The researchers found that focusing on CGT yielded substantial results, confirming its significance as a determining factor. In addition to labor productivity, the study also favored area productivity, measuring CGT against the total shipyard area, as suggested by Zhangpeng and Flynn (2006). Pires Jr et al. (2009) also explains that the most frequently used shipbuilding labor productivity indicator is labor hour per CGT.

Moyst and Das (2005), on the other hand, investigated factors affecting ship design and construction lead time and cost, which often incurs delays and exceed budgets. They found that situational factors, the industrial environment, and the overlap of engineering and production phases could impact lead times (Moyst and Das 2005). In another noteworthy study, Lamb and Hellesoy (2002) proposed a shipbuilding productivity predictor based on various shipyard characteristics. Their research revealed a high correlation between man-hours and CGT, suggesting its importance in predicting labor hours. Lastly, Semini, Brett, J. O. Strandhagen et al. (2022) analyzed various factors, concluding that ship size, complexity, building strategy, and global Offshore Support Vessel (OSV) demand influence the lead time of a vessel. These studies collectively provide a comprehensive understanding of the critical elements affecting labor hours in shipbuilding.

Given the high correlation, factor importance scores, and supporting literature on Compensated Gross Tonnage (CGT), it is deemed the most significant factor for this study. Despite not scoring high on correlation, the building strategy will also be included, following the work of Semini, Brett, J. O. Strandhagen et al. (2022). Notably, while there's scant evidence in the existing literature to suggest that the Series of the vessel significantly influences labor hours calculation, it will still be considered in this study. This inclusion is based on recommendations from industry experts and representatives from the case company under study, thus ensuring a comprehensive exploration of all potentially relevant factors

The following factors will be taken into account in further research:

- CGT
- Series
- Building strategy
- Outfitweight
- Vessel category
- Design
- Powerplant index

6.4 Model development

In the development of machine learning models for the current study, a systematic and robust approach was adopted to ensure accurate predictions and generalizability of the results. Cross-validation, specifically a 5-fold cross-validation, was utilized as a vital part of this approach. With this method, the dataset was divided into five subsets or 'folds'. During each iteration of the validation process, four folds were used for model training, and the remaining fold served as the test set. This approach not only provided a broader base of data scenarios for model training but also enhanced the evaluation's robustness by averaging model performance across the five folds (developers 2021). This is done because it provides a robust estimate of the model's performance on unseen data.

Importantly, this technique of cross-validation mitigates the risk of overfitting, ensuring the model's ability to generalize to unseen data effectively. The partitioning of data into folds was performed randomly using the `random-state` parameter set to a fixed integer, guaranteeing the reproducibility of the results across different runs. Thus, this 5-fold cross-validation method offers a more rigorous and reliable assessment of model performance in comparison to a simple train-test split.

The error metrics selected for this study, namely the coefficient of determination (R^2) and the Mean Absolute Error (MAE), have been chosen for specific reasons. The R^2 metric was primarily chosen due to the preference of the case company, but its usefulness extends beyond this. R^2 is a robust metric that gives a holistic view of the model's performance, assessing the proportion of variance in the dependent variable that is predictable from the independent variable(s). It provides an easily interpretable measure of how well the predictions approximate the real data points. In addition to R^2, MAE was used as a secondary metric. This decision was based on its ability to provide a direct interpretation of how far off the predictions are on average, making it a more intuitive measure of model performance. The MAE is particularly useful when the data contain outliers that can make other error metrics, such as MSE, disproportionately large. The combination of these two metrics allows for a comprehensive evaluation of the model's accuracy and predictive capabilities.

To construct the model, Python programming language was utilized along with various libraries such as NumPy, pandas, and Scikit-learn (Virtanen et al. 2020, Van Der Walt et al. 2011, McKinney 2010). The pandas library was particularly instrumental in loading and reading the historical data pertinent to all the models. Following data preparation and pre-processing, the construction of models was initiated, leveraging functionalities inherent in Scikit-learn. These performance metrics, R^2 and MAE, were calculated using the 'mean-absolute-error' and 'r2-score' functions, respectively, from the 'sklearn.metrics' module. The calculated values were derived from the average results on the test folds.

Furthermore, to ensure the robustness and reliability of the models, the strategy of cross-validation was incorporated into the modeling process. The datasets were divided into subsets or "folds" in a random manner, utilizing Python's "RandomState". Further, the data was segmented into training and testing sets through the application of cross-validation with a specification of five folds, employing the 'cross-val-score' function from the 'sklearn.model-selection' module. All the python scripts developed for each model can be found in the appendix.

6.4.1 Rationale Behind the Selected Prediction Models

In this study, the focus is on predicting the labor hours required for a project on a vessel—an area that is relatively less explored in scientific literature, especially in the context of employing Multivariate analysis models. Given the unique challenges associated with Engineer-To-Order (ETO) environments, particularly small datasets and low levels of industry digitalization, this field of study requires innovative approaches.

Framing our problem within the context of lead time prediction can help circumvent these challenges. Lead time prediction— the estimation of the time span from the inception to the completion of a process— shares key similarities with our focus on predicting labor hours for vessel projects. Extensive research in lead time prediction across various production environments, as noted earlier, provides a robust foundation to base our study upon.

The factors that influence lead time, such as resource availability, task complexity, and external influences, have significant overlap with the determinants of labor hours on a vessel project. Hence, there is a compelling rationale to adapt lead time prediction methodologies to our specific problem of labor hour prediction.

By extrapolating from the wider field of lead time prediction, we are able to bridge the gap caused by the limited research in labor hour prediction on vessel projects. This strategy provides us with a well-established, research-backed methodology, ensuring its appropriateness and applicability to the unique challenges of labor hour prediction in the less digitalized ETO industry context.

Litrature background

The selection of Machine Learning (ML) models for predictive analysis in diverse production environments has been extensively explored in the literature. A common application of these models is the prediction of lead time in semiconductor manufacturing. Lingitz et al. (2018) examined a multitude of ML models for this task, including Random Forest, Linear Regression, k-Nearest Neighbors (kNN), and Artificial Neural Networks (ANN), finding that Random Forest outperformed the others.

Similarly, Random Forest was found to be the best-performing model in an Engineer-To-Order (ETO) environment for predicting lead time in industrial and energy equipment manufacturing Kozjek et al. (2018). These findings align with the results of R. Murphy et al. (2019), who also found Random Forest to outperform other ML models in predicting flowtime.

The superiority of Random Forest was further validated by Pfeiffer et al. (2016), who found it to be the best model for lead time prediction in an Assemble-To-Order (ATO) environment. It appears that, across a range of manufacturing contexts, Random Forest consistently performs well.

However, in a Make-To-Order (MTO) environment, Neural Networks have been found to be the most effective model for similar tasks. This has been demonstrated by Schneckenreither et al. (2021), Su and Sha (2004), and Gacek (2018). These studies add a layer of nuance to our understanding of ML model performance, indicating that the optimal model may vary depending on the specific manufacturing environment.

Despite the success of Random Forest and Neural Networks, almost all cited studies have also applied Linear Regression methods for either explanatory modeling or as a performance baseline. This suggests the continued relevance of more traditional statistical techniques, underlining the importance of a comprehensive approach to model selection.

Justification of models developed

Multivariate analysis models are versatile tools in the domain of statistical analysis, providing capabilities for both explanation and prediction and often concurrently. Their explanatory power stems from their ability to delineate relationships among multiple variables, revealing complex patterns and underlying structures within the data. On the predictive front, these models are instrumental in forecasting outcomes based on input variables. Intriguingly, these two aspects often intertwine. An explanatory model that successfully uncovers the mechanisms of a system can serve as a solid foundation for accurate predictions. Conversely, a well-calibrated predictive model can provide insights into the relationships among variables, indirectly contributing to explanation. Thus, multivariate analysis models, with their dual capabilities, are critical tools for statistical exploration and inference, bridging the gap between understanding phenomena and anticipating future occurrences (Hair 2009). The case study shipyard wants a predictive model and, therefore, the main focus. However, among models with comparable performance, those with simpler interpretations are preferred.

Linear regression, despite its simplicity, remains a compelling choice for initial analyses, particularly for small datasets such as the one in consideration. As a low-complexity machine learning model, linear regression offers computational efficiency and interpretability, which is paramount when dealing with datasets where a transparent understanding of factor influences is desired (Draper and Smith 1998). This transparency can help in understanding the relationship between input parameters and hours needed for construction can lead to more efficient project planning. Furthermore, linear regression serves as an appropriate benchmark against which the performance of other, potentially more complex models can be compared. This is beneficial in the iterative process of model selection, ensuring we do not unnecessarily complicate our model if simpler alternatives suffice. Lastly, the choice of linear regression is backed by a wealth of literature endorsing its efficacy in similar contexts (R. Murphy et al. 2019), (Pfeiffer et al. 2016), (Mezzogori et al. 2019) . Therefore, the use of linear regression can provide a reliable starting point for our predictive modeling task and a solid foundation for any further refinements or comparisons.

Random Forest, an ensemble learning method, has consistently shown strong performance across a range of diverse datasets, including small ones, as in our case. Its inherent ability to handle non-linear relationships and interactions between factors makes it a powerful, versatile tool for prediction tasks (Breiman 2001). Unlike linear regression, Random Forest does not assume a particular type of relationship between the predictors and the response variable, providing a more flexible modeling approach. Additionally, Random Forest's robustness against overfitting, facilitated by the ensemble method's averaging mechanism, makes it an ideal choice when the aim is to generalize the model to unseen data (Cutler et al. 2007). Importantly, the Random Forest's performance is reinforced by a significant body of literature (Lingitz et al. 2018), (Kozjek et al. 2018), (R. Murphy et al. 2019). Numerous empirical studies across various domains have crowned Random Forest as the best-performing model, providing empirical evidence of its sufficient predictive power.

Semini, Brett, J. O. Strandhagen et al. (2022) finds through regression analysis on vessels built by Norwegian shipyards that Strategy, repeat production, size, complexity, and global OSV demand affect the lead time of a vessel. With this information, it would be logical to have models that cluster vessels with the same strategy together and the same level of repeat production. The factors of complexity and size are demonstrated to be the most correlated and significant contributors to the number of hours needed. This finding corroborates the factor importance calculations conducted in this study, where CGT achieved the highest score. Global OSV demand is an unavailable factor at this stage. For that reason, the cluster-then-regress method is used as it can cluster the categorical "Series" and "Building strategy" and linear regress with CGT. For clustering, Kmeans++ is chosen as the clustering algorithm as its widely used, low in complexity, and does not assume that each cluster needs to be the same size.

In pursuit of developing an accurate model for predicting working hours needed on a vessel, expert advice was taken into consideration to explore the significance of vessel category and design in our predictions. Vessel category and design can provide essential insights into the complexity, size, and functionality of the vessel, which can directly influence the working hours required to complete the project. Experts in the field recognize the importance of these factors, as they impact the overall efficiency and resource allocation in vessel operations. By incorporating vessel category and design into the model, it aims to capture the nuanced variations between different types of vessels and leverage this information to improve the accuracy and reliability of its predictions. Therefore, a two-step modeling approach is proposed: initially, a decision tree model is used to identify any similar vessels. Subsequently, a linear regression model is applied to this subset of similar vessels to predict the hours required for the new project.

Despite the literature suggesting that neural networks can be effective in similar fields (Schneckenreither et al. 2021), (Su and Sha 2004), and (Gacek 2018), it opted not to utilize them in this study for a couple of reasons. Firstly, neural networks typically require large datasets to perform optimally, while the dataset available for our problem is relatively small (Schmidhuber 2015). Using neural networks with limited data could lead to overfitting, resulting in poor generalization to new data points. Secondly, neural networks operate as black boxes, making it difficult to interpret and understand the reasoning behind their output (Dayhoff and DeLeo 2001). This lack of interpretability may hinder our ability to validate the model's predictions or draw insights that can inform decision-making in practical settings. Considering these limitations, it was decided to explore alternative models that are more suited to the dataset size and provide more interpretable results.

6.4.2 Explanation of developed models

The models have been justified, considering their performance and suitability for the given problem. It will now be provided a detailed explanation of each model, highlighting their intricacies, and present the comprehensive results obtained from their implementation. The results will be further presented in the section 6.6.1.

Non-linear regression model

During the model development phase, one of the challenges that were encountered was the relatively small size of the dataset. This presented limitations in terms of the complexity of models that could be employed, and there was a risk of overfitting if overly complex models were used. The first attempt by the yard was to use non-linear regression analysis on different clusters of building strategies to develop a model for the data. As shown in Figure 19, the resulting curve appeared to fit the data relatively well. However, upon further examination, it was determined that the non-linear regression approach was overfitted, as can be seen by the number of hours going toward zero when CGT increases. Given the small size of the dataset, this was not entirely unexpected.

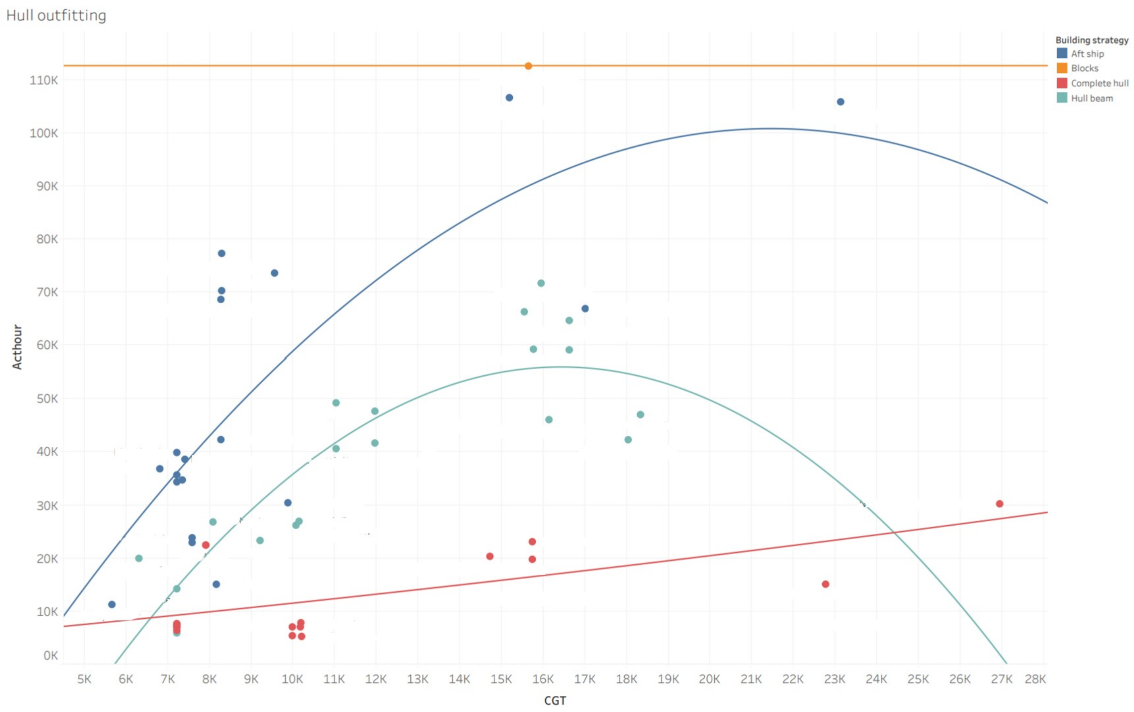


Figure 19: Non-linear regression on total hours

Linear regression model

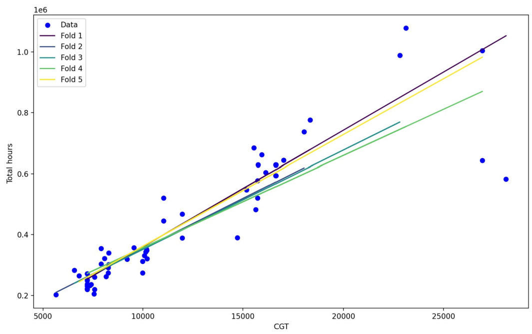
The issue of overfitting in the non-linear regression models is addressed by reducing model complexity. Overfitting often occurs when a model is too complex, capturing not only the underlying patterns in the data but also the noise, leading to poor generalization performance on new data (Hastie et al. 2009). To mitigate this issue, two different linear regression model variations were experimented with.

The first variation employed a simple linear regression, focusing on the most significant factor, CGT. This approach aimed to establish a direct relationship between CGT and the target variables. In contrast, the second approach utilized a linear regression model with multiple variables, considering the influence of several factors on the target variable. This more comprehensive approach sought to capture the combined impact of various factors, leading to a more nuanced understanding of the data.

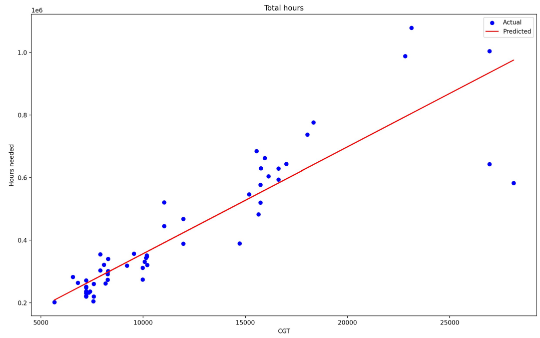
Both model variations are evaluated using 5-fold cross-validation. This approach provides a robust and reliable estimate of model performance by efficiently utilizing the available data, enabling the training and validation process to occur over multiple iterations (developers 2021).

In linear first approach it focused on building a regression model to investigate the relationship between the factor "CGT" and the number of hours needed. To achieve this, a linear regression model was employed that uses only the CGT factor as the predictor variable. The model is implemented using the 'LinearRegression' class from the 'sklearn.linear.model' module. The linear regression model is then fit to the training data using the 'fit' method, and the R^2 and MAE score is calculated on the average of the testing sets.

The results obtained from the analysis reveal that the linear regression model, when trained and tested using cross-validation, yields an $\mathbf{R^2}$ score of 0.656, indicating that the model explains approximately 65.5% of the variance in the number of hours needed based on the CGT factor. However, when the model is trained and tested without any train and test split, the $\mathbf{R^2}$ score increases to 0.8020. This was done to see the full linear relationship beetween CGT and total hours used on vessels.



(a) Linear regression with cross validation



(b) Linear regression without split

The linear second approach is building a linear regression model of the form: a * Series + b * Strategy + c * Outfitweight + d * CGT + e * Powerplant-index + X, which can provide insight into the relationship between various factors and a target variable (Montgomery et al. 2012). The benefits of using such a model include its simplicity, interpretability, and the ability to identify the impact of individual factors on the target. However, potential downsides include the assumption of linearity and the risk of overfitting or underfitting the data (Hastie et al. 2009).

The result of the cross-validation split approach got an R^2 value of 0.766 for the prediction of total hours. The formula for the first fold in the cross-validation was as follows:

$$-29127 * series + 13501 * Strategy - 60 * Outfitweight + 27 * CGT + 292 * Powerplantindex + 582$$

The result of this approach was The optimized R^2 value was 0.90 with the formula given under. The results of the optimization of the formula were as follows:

$$-26489 * SERIES - 5703 * Strategy + 99 * Outfitweight + 18 * CGT + 3984 * Powerplantindex + 23$$

Upon reflecting on the results, it is evident that the linear regression model was able to capture a significant portion of the variance in the data, as indicated by the relatively high R-squared value. However, it is crucial to consider the assumptions and limitations of linear regression models when interpreting these results (Hastie et al. 2009). The performace for each department are shown below.

Department	R^2	MAE
Total hours	0.656	65521
Outfitting yard	0.589	52955
Management	0.405	5044
Scaffolding	0.407	3641
Piping & machinery	0.758	7433
Production outfitting	0.766	12740
Engineering	-0.676	17133

Table 7: Linear vesion 1 predictive performance for each Departments

Department	R^2	MAE
Total hours	0.766	51906
Outfitting yard	0.718	47168
Management	0.471	4919
Scaffolding	0.642	2744
Piping & machinery	0.745	8579
Production outfitting	0.751	13295
Engineering	-0.361	16733

Table 8: Linear version 2 predictive performance for each Departments

The table below provides an overview of the results from our multivariate regression analysis, which was performed to understand the key drivers of the "Total hours" dependent variable in our dataset. We present the unstandardized and standardized coefficients along with corresponding p-values for each predictor or independent variable. The unstandardized regression coefficients (notated as β) represent the change in the dependent variable corresponding to a one-unit change in the respective predictor, assuming all other predictors are held constant. The standardized coefficients (notated as B), also known as beta coefficients, are the regression coefficients obtained when the variables in the model have been standardized to have a mean of zero and a standard deviation of one. They provide a means for direct comparison of the relative impact of each independent variable on the dependent variable. Lastly, the p-values help us assess the statistical significance of each predictor, with values below 0.05 typically considered as indicating a statistically significant relationship (Maneejuk and Yamaka 2021).

Parameter	β (Unstandardized)	B (Standardized)	p-value
constant (x)	55046.66	0.0000	0.3756
SERIES	-30137.32	-0.1338	0.0057
Building strategy	-4030.40	-0.0156	0.7539
Outfitweight	90.94	0.5120	0.0000
CGT	17.79	0.4669	0.0000
Powerplant_index	1909.73	0.0463	0.4921

Table 9: Regression Coefficients and p-values

Unstandardized coefficients represent the change in the dependent variable for a one-unit change in the predictor, keeping all other predictors constant. For example, for every unit increase in the Series, the Total Hours decrease by 30,137 hours, holding all other factors constant. Similarly, for every unit increase in Outfitweight, the Total Hours increase by approximately 90.94 units. On the other hand, standardized coefficients, often denoted as Beta or 'B', allow for comparison of the strength of the impact of each independent variable on the dependent variable. For instance, Outfitweight (B = 0.5120) and CGT (B = 0.4669) seem to have the most significant standardized impact on the Total Hours.

The p-values in the analysis provide insight into the statistical significance of the independent variables. The variables Series, Outfitweight, and CGT have p-values less than 0.05, commonly accepted as a threshold for statistical significance (Alexopoulos 2010). This means these variables significantly impact the Total Hours. However, the variables Building strategy and Powerplant_index have p-values higher than 0.05, suggesting they may not have a significant impact on the Total Hours.

Random forest model

The Random forest model ensemble method was applied to build a predictive model using factors CGT, Series, Building strategy, Outfitweight, and Powerplant index. Random forests are an ensemble learning technique that constructs multiple decision trees during the training process and combines their results to make a final prediction(Breiman 2001). This approach can provide a more accurate and robust model by reducing overfitting and increasing the model's generalization capabilities (Liaw and Wiener 2002). Moreover, random forests can handle non-linear relationships and automatically perform factor selection, which can offer insights into the importance of each factor in the model.

Random Forest, as an ensemble learning method, operates by creating a multitude of decision trees, each trained on different subsets of the data, and then amalgamating their outputs. The subsets are created through a process known as bootstrap sampling, which involves randomly sampling the original dataset with replacement to generate multiple distinct, yet overlapping subsets. This inherently diverse set of decision trees contributes to the robustness of the Random Forest model and its resilience to overfitting (Inc. n.d.).

When predicting, each decision tree within the Random Forest casts a vote for the outcome, and the final prediction is determined by majority voting, or in the case of regression, an average of the outcomes. This collective decision-making process helps mitigate the impact of any single decision tree's bias or variance, thus enhancing the model's overall predictive performance (Breiman 2001).

The random forest approach produced a model with an R^2 value of 0.820 for total hours, indicating that it was able to capture a significant amount of variance in the data. However, there is still room for improvement in the model's predictive accuracy.

The output of a random forest model is also the factors importance given in gini index. The analysis results strongly suggest that the 'CGT' factor is the most significant across most departments. In the model for 'Total hours', 'CGT' has an importance of approximately 64%, surpassing all other factors. This pattern remains consistent in 'Outfitting yard', 'Scaffolding', 'Production outfitting', and 'Piping & machinery', where 'CGT' holds the highest factor importance, ranging from 67% to 82%. This highlights the crucial role that 'CGT' plays in these departments.

However, in 'Project management' and 'Engineering', while 'CGT' still has significant importance, other factors notably 'Outfitweight' for 'Project management' and 'Series' for 'Engineering' show higher relevance. In estimation of total hours the factor importance was as follows: CGT (0.645), outfitweight (0.313), powerplant index (0.030), and Strategy (0.007) Series (0.006).

Department	$\mathbf{R^2}$	**MAE**
Total hours	0.820	50166
Outfitting yard	0.613	46100
Management	0.534	3864
Scaffolding	0.663	3045
Piping & machinery	0.786	7676
Production outfitting	0.752	11634
Engineering	0.181	9008

Table 10: Random forest predictive performance for each departments

Cluster-then-regress

Two versions of this method were attempted with version 1 presenting 2 different results as it was run twice. Both are utilizing Kmeans++ as the clustering algorithm.

KMeans++ algorithm was chosen due to its ability to partition data into distinct groups based on factor similarities, thereby revealing patterns within the dataset that may not be apparent when considering the entire dataset (MacQueen 1967a). Following the application of KMeans++, linear regression, a widely-used statistical method (Neter et al. 1996), was applied to the CGT factor within each cluster to predict the target variable. Although other clustering algorithms could potentially be suitable, given that KMeans++ fulfills the necessary criteria, further exploration of alternative methods was deemed unnecessary for this specific study.

For both versions the dataset was first partitioned into clusters based on the categorical values Series and Strategy, using the KMeans++ clustering algorithm (MacQueen 1967a). Subsequently, linear regression was performed within each cluster, utilizing the most important factor, CGT, to predict the target variable - the total hours needed for the project.

Version 1 - C-T-R1

There were trained 2 different types of C-T-R models in this case study. The first is only clustering on 1 variable at a time. Either building strategy or Series. The C-T-R1 (B) represents the model that clusters only on building strategy and then linear regresses on different strategies. C-T-R (S) represents the model that clusters on the factor Series and linear regresses in these clusters.

This approach first clusters the training data into clusters based on either factor "Building strategy" or "Series". When you use the k-means++ clustering algorithm and set the number of clusters to three, you will end up with three separate groupings. If you consider "Series" and "Building Strategy" as your variables, both having a range of values from 1 to 3, the same values will be grouped together. Afterward, it will do linear regression on these clusters and predict the hours for the department based on these regressions. Clustering with both "Series" and "Building strategy" individually was attempted, and the two tables below are showing the error metrics for both attempts.

Version 2 - C-T-R2

The second version clusters, with 3 clusters, on both "Series" and "Building strategy" at the same time in an attempt to find patterns yet not discovered.

Department	R^2	MAE
Total hours	0.272	26199
Outfitting yard	-2.833	48855
Management	-14.26	4179
Scaffolding	0.046	3517
Piping & machinery	-1.397	13682
Production outfitting	0.616	5820
Engineering	-6.847	5279

Table 11: C-T-R1 (S) predictive performance for each department

Department	R^2	MAE
Total hours	0.665	63495
Outfitting yard	0.777	27763
Management	-0.423	4187
Scaffolding	0.851	2407
Piping & machinery	0.665	63495
Production outfitting	0.661	6518
Engineering	-1.758	18329

Table 12: C-T-R1 (B) predictive performance for each department

The results demonstrate a reasonable degree of predictive accuracy, as reflected by the R^2 scores for the three clusters. However, it is essential to consider the potential limitations of this approach, such as the choice of clustering factors and the assumption of linearity within clusters.

Department	R^2	MAE
Total hours	0.828	59236
Outfitting yard	-1.275	96163
Management	-14.467	10343
Scaffolding	-0.558	6648
Piping & machinery	0.085	19077
Production outfitting	0.894	5491
Engineering	0.345	9099

Table 13: C-T-R2 predictive performance for each Departments

The results show inconsistent R^2 values in different departments. It was also inconsistent when looking into each cluster. For instance, the R^2 ranges from -2 to 0.97 across the different clusters for the same department. This indicates that the model's performance is inconsistent and may not be well-suited for predicting department-specific hour usage.

To address these limitations, the clustering step was modified to include only two clusters in an attempt to simplify the model and potentially improve its performance. However, this alteration did not yield any significant improvements in the performing metrics.

Sequentially Restricted Regression (SRR)

In the given approach for predicting labor hours in potential projects, historical data is utilized to make informed decisions based on the similarities in vessel categories and designs. The methodology takes into account the value of the "Series" variable, which indicates if the vessel is similar or alike to a previous project. If the Series is 1, meaning it is the first of its kind, a regression analysis is performed on all the vessels with Series=1 and their corresponding CGT to predict the required labor hours. This approach leverages the collective knowledge of previous first-time vessel projects to estimate the labor hours needed for this new project.

When the Series is 2, suggesting that the vessel is similar to a previous one, the algorithm narrows down its search and considers only those vessels within the same "Vessel category" as the new potential project. The average labor hours per CGT for these vessels are calculated and then multiplied by the CGT value of the example to predict the required labor hours. Similarly, if the Series is 3, meaning the vessel is essentially identical to a previous one, the algorithm becomes even more specific and focuses on vessels with the same "Design" as the example. The average labor hours per CGT for these vessels are then multiplied by the CGT value of the example to obtain the predicted labor hours.

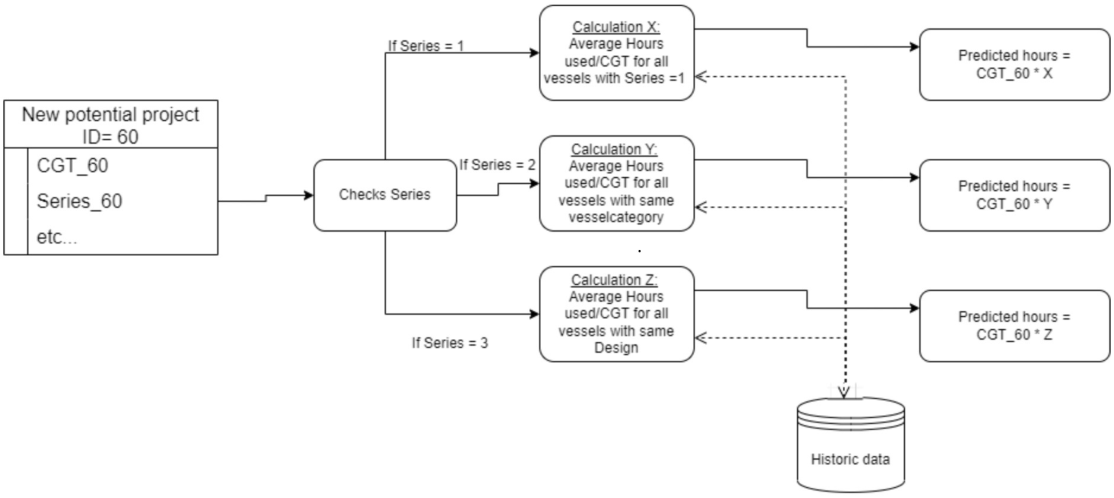


Figure 21: Visualization example of the Sequentially Restricted Regression model

The effectiveness of this approach is demonstrated by the R^2 score of 0.935 on total hours, which indicates a strong correlation between the predicted and actual labor hours. Given the nature of this model, which necessitates data in a sequential, chronological format, certain limitations in its applicability may surface. For instance, should a vessel share its design with only one other vessel, and that vessel is included in the training set, the model would attempt to find other vessels of the same design for learning purposes. However, in such a case, the script would run into an error. Consequently, owing to these model-specific intricacies, the conventional approach of cross-validation was not feasible. Instead, an alternative method of partitioning the dataset was adopted. This involved implementing an 80/20 train-test split, which better suited the chronological nature of the data and the model's specific requirements. However, to increase the model's reliability, different possible 80/20 splits were attempted without any significant differences in results being detected. The results in the tables are from the first random 80/20 split.

Department	$\mathbf{R^2}$	**MAE**
Total hours	0.935	25449
Outfitting yard	0.909	21401
Management	0.498	4287
Scaffolding	0.795	3210
Piping & machinery	0.035	12680
Production outfitting	0.933	4093
Engineering	0.5	11423

Table 14: SRR predictive performance for each department

6.5 Model validation

To ensure the reliability and accuracy of the developed predictive models, a comprehensive model validation process was conducted. In the process of model development, considerable adjustments have been made to enhance their performance and reliability. The models depicted above have been meticulously honed over numerous iterations, each designed to refine their predictive capabilities. This iterative development process was facilitated through the application of various model validation techniques. This section outlines the key aspects of the validation process, including data splitting, cross-validation, performance metrics, model comparison, hyperparameter tuning, and model robustness.

Data Splitting:
The dataset was divided into a training set and a testing set using an 80/20 split. This means that 80% of the data was used for training the models, while the remaining 20% was reserved for testing and evaluating their performance. This approach helps to prevent overfitting and ensures that the models are capable of generalizing well to new, unseen data. The first round of model development this was done alone, but after some strange results showing for some models the cross-validation was employed to help with validation of the results.

Cross-Validation:
In addition to data splitting, when feasible, the technique of cross-validation was integrated to bolster the assessment of the models' performance, thereby contributing to a more robust validation framework. K-fold cross-validation, with k=5, was implemented, which involves partitioning the training data into k equal-sized subsets and iteratively training the models on k-1 subsets while validating on the remaining subset. This process was repeated k times, and the average performance metric was calculated to obtain a more robust estimate of the models' performance.

Performance Metrics:
Various performance metrics were utilized to evaluate the models, including the mean absolute error (MAE) and the coefficient of determination (R^2). The use of these two error metrics was adopted to leverage their complementary characteristics, thus providing a more comprehensive evaluation of the model's performance. The MAE, by focusing on the average absolute difference between predicted and actual values, quantitatively captures the magnitude of error in the model's predictions. However, this metric alone can be abstract as it does not provide a relative measure of how well the model performs in the context of the problem space. Hence, the R^2 metric was also employed. R^2 presents the proportion of the variance in the dependent variable that is predictable from the independent variables, thus providing an interpretable measure of the model's predictive power in relation to the total variability in the data. In this way, the combination of MAE and R^2 offers a balanced and holistic measure of the model's performance.

Model Comparison:
The performance of the different models was compared based on the calculated performance metrics. This comparison allowed for the identification of the most effective model and provided insights into their strengths and weaknesses. The comparison of results are presented in section 6.6.1 and the discussion of the results can be found in 7.2

Hyperparameter tuning:

Hyperparameter tuning is a critical step in machine learning model development. It involves adjusting the configurations of the model to optimize its performance. In this project, hyperparameters were meticulously adjusted across various methods. For example, in the context of cross-validation, the number of folds (k) was varied to find the optimum balance between bias and variance, thus enhancing the model's generalizability. In the case of the Kmeans++ clustering algorithm, the number of clusters was adjusted. Identifying the optimal number of clusters is pivotal as it directly impacts the effectiveness of the clustering process and the resultant insights drawn from the data.

Likewise, in the Random Forest model, hyperparameters such as the number of decision trees in the forest (n-estimators), and the maximum depth of the tree (max-depth) were fine-tuned. The number of decision trees can influence the model's ability to capture complex patterns in the data, while the maximum depth can control the model's complexity, preventing overfitting. Through rigorous hyperparameter tuning, we aimed to optimize each model's predictive performance while maintaining their robustness and interpretability.

Model Robustness:
The robustness of the models was assessed by introducing small perturbations in the input data and analyzing their impact on the model's predictions. This analysis helped to identify any potential issues related to the models' sensitivity to small changes in the input data and provided insights into their overall stability and reliability. At the start of the development, there was an error connected to missing values. This problem was solved by adding code to drop rows that were missing necessary calculations. With this solution, there can occur a problem with the SRR model. These cases will be if an equal vessel is being built, but the information about the previous vessel is deleted due to missing values. This is a potential problem and did not occur during the testing period.

6.6 Results and evaluation of models

This subsection will present the results from all models and a comparison among them. There will be discussed which of the models performs the best on different departments. possible downsides and benefits with each model will also be discussed

6.6.1 Results

In this section, we present the results of a case study focused on predicting the number of hours needed in different departments of a shipyard, utilizing historical data. Six machine learning models were employed for this task: two linear models (Linear 1 and Linear 2), Random Forest, two cluster then regress models (C-T-R 1 and C-T-R 2), and the Sequentially restricted regression model. C-T-R 1 are shown with results with two different factors that are clustered. Once clustered with the factor "Series" and another with "Building strategy".

The performance of each model was checked using two measures: R^2 and the MAE for the hours needed in the selected departments. The evaluation results presented are derived from the testing part of the dataset in the cross-validation process, thereby providing a more robust and reliable measure of model performance. These findings represent an average of over five folds when they were included in the test set, reinforcing the validity and generalizability of the results. This was not possible for the Sequentially restricted regression model as it needs chronological data and is not at random. SRR is therefore utilizing an 80/20 test-train split.

The table below presents the comparison of different methods for each department based on the performing measure R^2:

Department	LR 1	LR2	RF	C-T-R 1 (S)	C-T-R 1 (B)	C-T-R 2	SRR
Total hours	0.656	0.766	0.820	0.272	0.665	0.828	0.935
Outfitting yard	0.589	0.718	0.613	-2.833	0.777	-1.275	0.909
Management	0.405	0.471	0.534	-14.26	-0.423	-14.467	0.498
Scaffolding	0.407	0.642	0.663	0.046	0.851	-0.558	0.795
P & M	0.758	0.745	0.786	-1.397	0.665	0.085	0.035
Prod outfitting	0.766	0.751	0.752	0.616	0.661	0.894	0.933
Engineering	-0.676	-0.361	0.181	-6.847	-1.758	0.345	0.5

Table 15: Comparison of Different Methods for Each Department - R^2

The Sequentially restricted regression model achieved the highest predictive accuracy for total hours, with a metric value of 0.934, followed by the C-T-R 2 model at 0.827 and RF at 0.820. For individual departments, the performance varied across different models. In the Outfitting yard department, the SRR model again demonstrated the highest predictive performance (0.909), while the C-T-R 2 and C-T-R 1 (S) showed poor performance with negative metric values. In the Management department, none of the models performed exceptionally well, with the highest metric value being 0.534 for the random forest model. The SRR model achieved relatively average performance in this department, with an R^2 value of 0.498.

In the Scaffolding department, the C-T-R 1 (B) outperformed the other models, with a metric value of 0.851. The Random Forest model achieved the best performance in the Piping and Machinery department (0.786). However, the SRR model showed poor performance in this department, with a metric value of 0.035.

In the Production Outfitting department, as illustrated in Table 15, the SRR model emerged as the most effective with an R^2 value of 0.933. This was closely trailed by the C-T-R 2 model, which registered an R^2 value of 0.894, and the Linear Regression 2 (LR2) model, with an R^2 value of 0.751. Notably, the C-T-R 1 (B) model showcased a comparatively moderate performance, as reflected in its R^2 value of 0.661.

The table below presents the comparison of different methods for each department based on the performing measure mean absolute error:

Department	LR 1	LR 2	RF	C-T-R 1 (S)	C-T-R 1 (B)	C-T-R 2	SRR
Total hours	65521	51906	50166	26199	63495	59236	25449
Outfitting yard	52955	47168	46100	48855	27763	96163	21401
Management	5044	4919	3864	4179	4187	10343	4287
Scaffolding	3641	2744	3045	3517	2407	6648	3210
P & M	7433	8579	7676	13682	63495	19077	12680
Prod outfitting	12740	13295	11634	5820	6518	5491	4093
Engineering	17133	16733	9008	5279	18329	9099	11423

Table 16: Comparison of Different Methods for Each Department - MAE

From the results presented in Table 15 and 16, the performance of the six machine learning models varies across the different departments in the shipyard. The Sequentially restricted regression model achieved the highest predictive accuracy for the total hours and demonstrated strong performance in several departments. However, it also showed poor performance in the Piping and Machinery department. The Random Forest model exhibited consistent performance across multiple departments, while the C-T-R models showed mixed results, with both strong and weak performance in different departments.

These findings suggest that no single model outperforms others in all the departments consistently. Instead, each model demonstrates strengths and weaknesses in different departments. Further analysis and fine-tuning of the models may help in improving their predictive accuracy across all departments. Additionally, exploring ensemble techniques to combine the strengths of multiple models could be a viable approach to achieve better performance across all departments in predicting the hours needed in the shipyard.

To visualize the results better all the values smaller than 0 are set to 0 in figure 22 for better visualization.

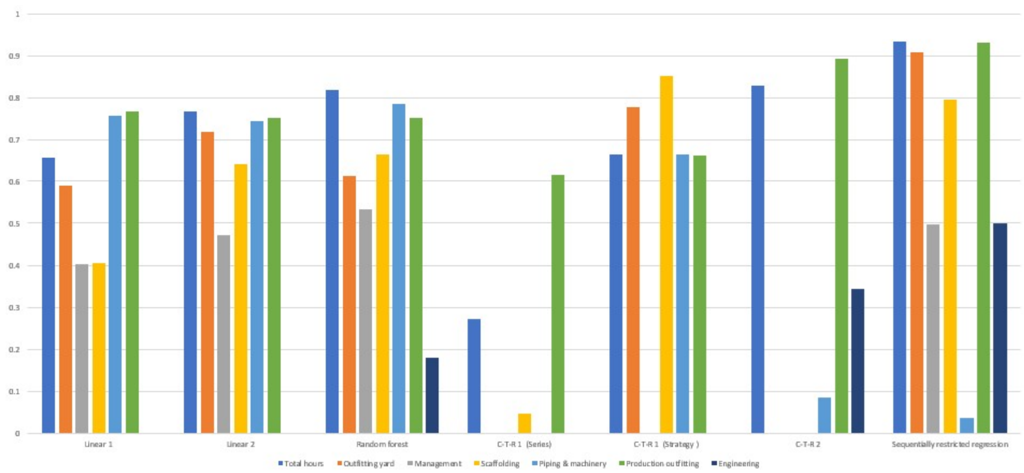
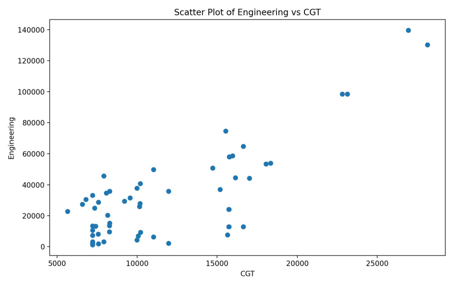


Figure 22: Visualization of the different models on departments - R^2

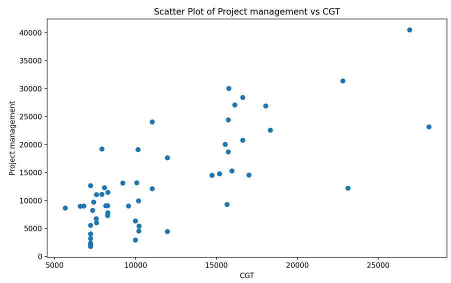
Figure 23: Visualization of the different models on departments - MAE

In evaluating the predictive models across various departments, the fields of Engineering and Management pose the most significant challenges. The relatively lower number of hours in these departments might be a contributing factor to this unpredictability. When assessing the MAE, it's fairly low compared to the other departments. More strikingly, the R^2 values indicate a more substantial discrepancy. For instance, in Management, the R^2 values ranged from -14 to 0.534 across different models, with the Random Forest model yielding the highest predictive value at 0.534. In contrast, the Engineering department, despite having an avaerage higher range of R^2 values, still underperformed compared to other departments, and the best model scored a R^2 value of 0.5. It is noteworthy that the C-T-R 1 models demonstrated negative R^2 values for both departments, but the C-T-R (S) is at the same time showing good performance based on MAE.

The difficulty in predicting the 'Engineering' and 'Management' categories can be elucidated further through the scatter plots presented below. Upon analysis, it is apparent that these categories do not exhibit the same level of discernible patterns compared to others such as 'Total Hours', as depicted in Figure 20b. This scarcity of predictable trends significantly enhances the complexity of forecasting these specific categories, thereby underscoring the challenge at hand.



(a) Scatterplot of hours in engineering department vs CGT



(b) Scatterplot of hours in management department vs CGT

6.6.2 Discussion and evaluation

One model might outperform others for a variety of reasons. Firstly, the superiority of one model could be due to the nature of the data. For instance, if the data has a complex, non-linear structure, a model such as Random Forest that can capture such complexity might outperform a linear regression-based model. From the case study, it was found that it was a linear relationship between CGT and total hours with an R^2 score of 0.8020. This means that by just looking at CGT, it's possible to explain 80% of the variability in total hours.

Secondly, the quality and type of factors can influence a model's performance. Some models are more robust to noisy or irrelevant factors. If the dataset includes many such factors, a robust model would likely outperform others. Lastly, overfitting could be a factor. Models like Random Forest have built-in mechanisms to prevent overfitting (Breiman 2001). If the models are trained on a small dataset or a dataset with many factors, models with such mechanisms might perform better.

A critical aspect of machine learning models that should not be overlooked is their interpretability. While Random Forests have proven to be a robust predictive tool in our study, they often serve as "black box" models, providing limited insight into how to input factors are being used to make predictions (Inc. n.d.). This is contrasted to the other models presented, which offer more transparency into the factor-to-prediction pathway. The trade-off between accuracy and interpretability is a crucial consideration. While a highly accurate model is desirable, it is equally important, especially in strategic decision-making contexts, to understand the 'why' behind a prediction. This underscores the need for balancing model complexity with interpretability in real-world applications.

Linear version 2 seems to be overfitting in some departments. This can be seen in engineering, where the average R^2 for the Training data was 0.609, but for the unseen test set, the average R^2 was -0.366. Can also be seen by the difference in formulas from the 5 folds on total hours prediction:

$-29127 * \text{SERIES} + 13502 * \text{Strategy} + 61 * \text{Outfitweight} + 28 * \text{CGT} + 292 * \text{Powerplant\_I} + 583$

$1 * \text{SERIES} + 1 * \text{Strategy} + 73 * \text{Outfitweight} + 22 * \text{CGT} + 0 * \text{Powerplant\_I} + 1$

$28 * \text{SERIES} - 37 * \text{Strategy} + 85 * \text{Outfitweight} + 21 * \text{CGT} - 115 * \text{Powerplant\_I} - 37$

$1 * \text{SERIES} + 1 * \text{Strategy} + 73 * \text{Outfitweight} + 22 * \text{CGT} + 0 * \text{Powerplant\_I} + 1$

$1 * \text{SERIES} + 1 * \text{Strategy} + 70 * \text{Outfitweight} + 22 * \text{CGT} + 0 * \text{Powerplant\_I} + 1$

The varied prediction formulas we see in the output above seem to be an indication of a potentially overfitted model. In this case, the model seems to be generating a multitude of formulas for prediction, which suggests it's likely over-adapting to the specifics of the training data, thereby reducing its predictive power on unseen data. The complexity of these formulas, along with their varying coefficients, point towards a model that might be fitting too closely to the noise or outliers in the training data rather than identifying the underlying trend. The model's performance metrics also suggest possible overfitting. We observe a relatively high Mean Absolute Error (MAE) for the test set as compared to the training set, indicating that the model isn't generalizing well on unseen data. Can also see from table 9 that the p-values for "Building strategy" and "Powerplant-index" is at a level of 0.75 and 0.49. These high p-values signify that the possibility of obtaining the observed data, given that the null hypothesis is true, is fairly high. Therefore, it's less likely that the observed effect in "Building strategy" and "Powerplant-index" is due to any real effect; rather, it might be attributed to random chance. The model will, therefore, in this case, be seen as to complex for this dataset.

The SRR model, which exhibited the highest predictive accuracy (R^2) for total hours, demonstrated strong performance in several departments, including the Outfitting yard, Production outfitting, and Engineering. However, it also showed poor performance in the Piping and Machinery department. This suggests that the SRR model may be well-suited for certain departments but may not be the best choice for others. Its also important to remember that the model's evaluation lacks the robustness provided by cross-validation, which could not be applied in this case. As such, the model's performance indicators might not be as reliable or generalizable across different data sets. However different 80/20 splits was attempted to check the reliability of the model, and no big differences was detected.

The SRR model's performance in the various departments of the shipyard raises interesting points for discussion. It is noteworthy that the model achieves high scores in the Outfitting yard, Production outfitting, and total hours, but not in Management, Piping and Machinery, and Engineering. One possible explanation for this discrepancy could be the varying impact of the vessel category and design on different departments. The Outfitting yard, Production outfitting, and Total outfitting hours might be more influenced by the this, as the construction processes and requirements in these departments could be more similar for vessels built in the same category. In contrast, the Management and Engineering departments may not be as affected by the vessel category factor. This can also be seen in the factor importance from random forest where managment and engineering showed the factor "Series" to be the most important factor.

It is somewhat surprising that the SRR model does not perform better in the Piping and Machinery department, given that similar vessels are likely to have comparable systems integrated. A possible reason for this observation could be that the SRR model might not adequately capture the intricacies and dependencies of the Piping and Machinery department. It is also possible that other factors not considered in the model may influence the number of hours required in this department, causing the model's performance to decline.

Further investigation into the underlying reasons for the SRR model's performance in the Piping and Machinery department could help improve the model's predictive accuracy and applicability across all departments.

When looking at the table 18, it can be seen that "Series" have a strong correlation with engineering and management, but the SRR model will cluster the vessels based on vessel category and design based on the factor "Series" and not the series itself. This makes the way of the C-T-R1 (S) an interesting model. It outperformed other models by achieving the lowest MAE scores for both the engineering and management departments. This indicates that this method was effective in minimizing the average magnitude of prediction errors for these departments.

However, this method also demonstrated certain limitations, notably in terms of the R^2 score. For engineering and management departments, this model yielded the worst and second worst R^2 scores among the models tested. The low R^2 scores suggest that, despite having a lower average error magnitude, the model may not be capturing the overall variation in the data very well.

Interestingly, factor importance analysis from the random forest model supported the relevance of the "Series" factor in these departments. The "Series" factor emerged as the most important variable for predicting labor hours in engineering and management departments. This suggests that the cluster-then-regress method, which explicitly incorporates this factor through its clustering step, has a sound underlying rationale. Figure 26 visualizes the connection between Series and labor hours in engineering.

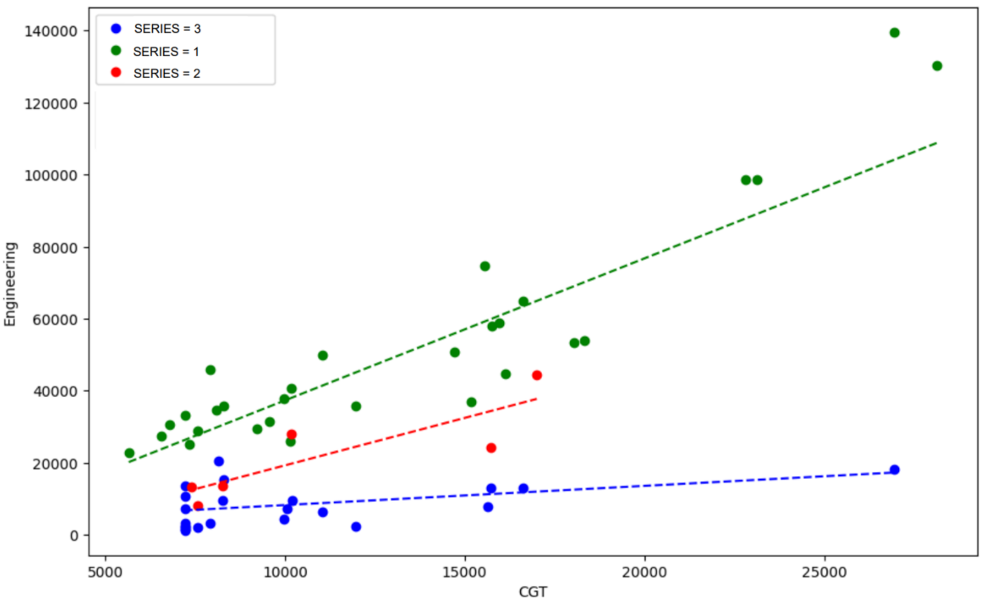


Figure 25: Visualization of C-T-R with 1 cluster

The lower R^2 score in the cluster than regress methods, as compared to standard linear regression, can be attributed to the fact that this method involves creating three separate regression lines (one for each cluster) instead of a single line. This process inherently introduces more complexity into the model, which can lead to a lower proportion of the variability in the dependent variable being captured by the model, thus resulting in lower R^2 scores, but still a low MAE.

The Random Forest model demonstrated consistent performance across multiple departments, including Piping & Machinery, and management, where it outperformed the other models. This suggests that the Random Forest model may be a more reliable choice for these departments. However, it may not be the optimal choice for other departments where other models have demonstrated better performance.

The historical data used in this study is relatively small in size, which has implications for the performance of the machine learning models used. Generally, machine learning algorithms perform better when trained on larger datasets, as they have more information to learn from and can capture the variability and complexity of the data more accurately. Therefore, it is reasonable to anticipate that the performance of all the models would improve with an increase in the size of the data.

Moreover, the implementation of the models developed in this thesis will lead to a 25% increase in the size of the dataset as it's trained on the whole dataset and has no 80/20 split. This will happen naturally as new projects are initiated and completed, with data from each project added to the existing dataset. This continuous influx of new data will not only improve the accuracy of the models but also help in capturing any evolving trends or changes in shipyard operations.

Therefore, it can be expected that the actual performance of the models, when implemented in real-world scenarios, will be even better than what is reported in this thesis. This makes the findings of this study not only promising but also scalable and adaptable to the dynamic nature of shipyard operations.

A noteworthy outcome derived from the case study was the statistical significance of the factors. In alignment with prior literature of similar nature, CGT was identified as the most important factor through person correlation and the gini index derived from random forest. When linear regression was applied solely on CGT against total hours, a coefficient of determination (R^2) score of 0.802 was observed, emphasizing its notable predictive capability. However, the consequential influence of repetitive practices, as represented by the "Series" factor, on the departments of Engineering and Project Management presents a unexplored area of exploration. Furthermore, the impact of "Vessel category" and "Design" were presented their performance in predicting total hours on unseen data, attaining an R^2 score of 0.935. Such findings reinforce the multifaceted nature of these factors and their extensive impact on the predictive model's performance.

6.6.3 Justification and recommendation of models for each department

Based on the findings, evaluation, and discussion, the recommendation is to implement different models for different departments to achieve optimal performance.

Total hours, Outfit yard, and Production outfitting

Sequentially Restricted Regression (SRR) stands out as a highly recommended method for predicting total hours, hours needed for the outfitting yard, and production outfitting. Demonstrating superior performance on key evaluation metrics. The method surpasses other models in terms of both R^2 and Mean Absolute Error (MAE). The higher R^2 score indicates that the SRR model accounts for a greater proportion of the variance in the dependent variable. In other words, it is more effective in capturing the overall trend in the data and provides a stronger fit to the data points.

The lower MAE, on the other hand, means that the SRR model is successful in minimizing the average of the absolute differences between the predicted and actual values, leading to fewer prediction errors. These strong performances on both metrics underline the model's predictive accuracy and reliability.

In light of these findings, the Sequentially Restricted Regression is an excellent choice for forecasting labor hours in these areas. Its superior performance in both capturing the data's variance and reducing prediction errors suggests that it will deliver reliable and accurate predictions, making it a valuable tool for managing labor hours within shipyard operations.

Engineering
The cluster then regress model, specifically using the "Series" factor for clustering, emerges as a strongly recommended approach for the engineering department within the shipyard operations. The engineering department, being one of the smaller departments and demonstrably the most challenging to predict in this study, requires a robust prediction model.

Despite presenting lower R^2 values, the cluster then regress model significantly outperforms other methods on the Mean Absolute Error (MAE) metric. This measure reflects the average magnitude of prediction errors, irrespective of their direction, and thus provides a practical gauge of predictive accuracy. The closest competitor, the Random Forest model, exhibits an MAE that is 70% higher, underscoring the superior accuracy of the cluster-then-regress model in this context.

Given the criticality of labor costs in Norwegian shipyard operations, a model that minimizes prediction errors is of paramount importance. Therefore, with its strong performance on MAE, the cluster then regress model with "Series" clustering offers a promising strategy for better managing labor hours and costs within the engineering department.

Management
The Random Forest model emerges as the optimal choice for managing labor hours within the management department, given its demonstrated strengths in the case study. The management department, being among the smaller entities and ranking as the second most challenging to predict, requires a model capable of handling its intricate dynamics.

The Random Forest model delivers superior performance on both key metrics. A higher R^2 value signifies that the model explains a large portion of the variance in the dependent variable, indicating a robust fit to the data points. Simultaneously, a lower MAE suggests that the model successfully minimizes the average of the absolute differences between the predicted and actual values, thus reducing prediction errors.

Random Forest is known for its ability to model complex relationships due to its ensemble learning approach, which can explain its strong performance with the management department. It effectively captures the high complexity of the department's operations, enabling accurate prediction and thereby assisting in efficient labor hours management. Therefore, for the management department, the Random Forest model is recommended due to its reliable accuracy and robustness in capturing complex patterns.

Scaffolding

The cluster then regress model, specifically employing "Building Strategy" for clustering, is recommended for the scaffolding department due to its superior predictive performance. In terms of both R^2 and Mean Absolute Error (MAE), this model outperforms others, demonstrating both high explanatory power and accuracy in its predictions.

In the dataset pertaining to the scaffolding department, only the hours of scaffolding work performed at the case company's yard are registered, excluding the work completed abroad. It is, therefore, logical that fewer hours would be necessary when a greater proportion of work is performed in Eastern Europe. This is evident in the figure below, where it is apparent that strategy 3 yields a reduced scaffolding hours requirement. The difference between Strategies 1 and 2 is relatively insignificant, further indicating that Strategy 3 provides a distinct advantage in terms of reducing labor hours. The figure below is trained on the whole dataset and has an R^2 of 0.910 and MAE of 1749.

The cluster then regresses the model's ability to discern these strategic influences on labor hour requirements, thus making it an advantageous choice for the scaffolding department. By accurately reflecting the impact of different building strategies on labor hours, this model can support more effective planning and management within the department.

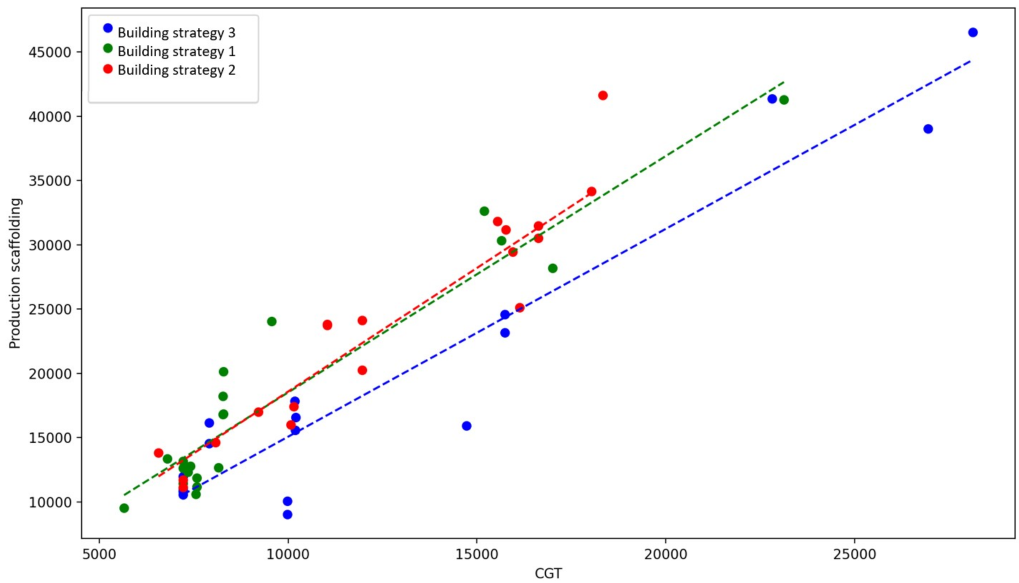


Figure 26: Visualisation of Clustered and Regressed Scaffolding by Strategy

Piping & machinery

Linear regression was selected as the preferred predictive model for the Piping and Machinery department based on several compelling reasons. Firstly, it displayed the second highest R^2 value, ranking just behind the Random Forest model, indicating a robust explanatory power for the variability in the data. More importantly, it yielded the lowest Mean Absolute Error (MAE), thus illustrating its superior accuracy in predictions. Additionally, the theoretical underpinning that piping and machinery are highly dependent on CGT aligns seamlessly with the linear regression model, which is adept at elucidating such relationships.

Despite the close performance between linear regression and Random Forest, the former was ultimately chosen due to its inherent strengths. Specifically, linear regression offers greater explanatory power, making it easier to understand and interpret the influence of predictor variables. Moreover, it possesses simplicity, making it a more convenient and efficient choice for implementation and use.

6.6.4 Summary of final results from the case study

The case study results highlight the importance of considering the strengths and weaknesses of different machine learning models when applied to various departments within a shipyard. Implementing different models for different departments or utilizing ensemble techniques may lead to improved performance in predicting the hours needed in the shipyard.

Alternatively, ensemble techniques could be explored to combine the strengths of multiple models and achieve better performance across all departments. Ensemble methods, such as stacking, bagging, or boosting, may help improve the overall performance by leveraging the diverse strengths of individual models. For instance, a weighted combination of the SRR and Random Forest models could potentially yield improved performance in departments where the SRR model underperforms, such as the Piping and Machinery department.

Further analysis and fine-tuning of the models may also help in improving their predictive accuracy across all departments. This may involve hyperparameter optimization, factor engineering, or the application of more advanced machine-learning techniques. Additionally, incorporating domain knowledge and expert insights in the development and evaluation of these models may enhance their applicability and relevance to the specific context of the shipyard.

In conclusion, the recommendations are as follows; use of SRR on Total Hours, outfit yard, and production outfitting. C-T-R with clusters on "Series" for engineering, Random forest to use on the management department, cluster then regress model, specifically employing "Building Strategy" for clustering, for scaffolding. The performance of all suggested models surpasses that of the basic linear regression executed on the entire dataset, which was regressed on CGT. This implies that these models encapsulate more than just a simple linear correlation inherent in the data, signifying a higher level of complexity in their predictive capabilities.

7 Discussion

This section will aim to answer the research questions presented in the introduction. It will discuss how the concept of a digital twin can be introduced and applied in Norwegian shipyards to improve their operations. It will also be discussed how digital twin technology influence decision-making processes in managing labor hours within shipyard operations.

7.1 Research question 1

RQ1: How can the concept of a digital twin be introduced and applied in Norwegian shipyards to improve their operations?

The concept of digital twins, a virtual representation of a physical system, holds great promise in enhancing the operational efficacy of several larger Norwegian shipyards (Kunkera et al. 2022). These representations can be leveraged to optimize various yard functions, including layout and logistics, resource and capacity management, as well as disaster preparedness and emergency response. Furthermore, the predictive maintenance potential of digital twins can minimize system downtime and augment the overall lifecycle of shipyard assets (Süve et al. 2022). However, successful deployment of this novel approach entails requirements such as investment in required technology, investment or fostering of necessary expertise, and availability of quality data.

While the benefits are evident, the feasibility of extensive investments in technology and expertise has become a challenge due to the downturn in vessel demand, specifically for Norwegian shipyards previously reliant on vessels for the oil and gas industry. The financial constraints have instilled a cautious approach towards large investments without demonstrable returns in the shipbuilding industry. Consequently, an incremental approach to digital twin implementation has been suggested. This method begins with the adoption of a few key applications, gradually integrating and building upon them over time, thus reducing the initial investment burden and allowing for tangible results before further advancements. This phased approach to digital twin integration allows shipyards to capitalize on the potential benefits while managing associated risks and costs.

Given the requirements, it appears plausible to initiate the integration of digital twins within the administrative operations of the shipyard. This assertion is supported by representatives from a shipyard who believe in the sufficiency of high-quality data within these operations. The case study has also demonstrated the viability of such an application, revealing that the data quality is sufficient to build high-performing digital twin-based applications. Given that industry experts and shipyard representatives perceive significant potential in applying the digital twin concept within resource management, it appears to be a prudent starting point for its implementation. These operations are crucial to the functioning and profitability of the shipyard, and any improvements in efficiency and accuracy can lead to significant cost savings and operational optimization.

The application of digital twins can be extended to functions like lead time prediction, demand forecasting, cost estimation, and process estimation. These applications, when embedded in the resource management function, can augment operational effectiveness, contributing to an overall enhancement in efficiency. Thus, the gradual integration of these technologies, starting with the digital twin concept, can potentially transform shipyard operations, setting a new paradigm in resource management and operational optimization.

The case study provided a concrete demonstration of the digital twin's predictive capabilities by accurately forecasting labor hours required in shipyard operations. It not only proved the practicality and effectiveness of the digital twin concept but also showed its potential to improve resource allocation and efficiency. This foundational application of labor hour prediction can be incrementally built upon, incorporating physical assets in the digital twin, like docks, which allows for a comprehensive simulation of resources, enabling better control over resource allocation and utilization. This application also fit together nicely with the concept of predictive maintenance. As proven possible through the use of the Internet of Things (IoT) and Machine Learning (ML) (Aivaliotis et al. 2019), predictive maintenance can ensure timely maintenance of equipment, reducing downtime and costs associated with equipment failure.

By merging applications as resource management and predicted maintenance, it should also be possible to better the decision-making around investment in new machinery. Hagen and Erikstad (2014) names investments like providing structural support for upgrading cranes, investing in hybrid powering systems, and additional power reserves as complex decisions. These decisions can be helped by information regarding the capacity and downtime of machinery in addition to reliable resource predictions. A functional digital twin for these applications should also be able to simulate different scenarios to help decision-making.

This demonstrates the advantages of integrating diverse applications within the digital twin framework to extend its utility across various domains. By not only harnessing the benefits from individual applications but also facilitating synergy amongst them, it's feasible to enhance the overall efficacy of the digital twin and unlock additional benefits. Such a collaborative approach to application integration accentuates the multifaceted advantages of the digital twin concept.

There are relatively few established applications of the digital twin concept within the shipbuilding industry at present, with most efforts concentrating on the vessel itself rather than the yard operations (Resyard 2023). This leaves open a compelling question: whether it is economically beneficial to position oneself as an industry leader in this emerging field. A noteworthy application of digital twin technology in the shipbuilding industry, identified through the literature review, was the use of 3D modeling of the shipyard. This innovative approach reportedly resulted in a productivity boost of approximately 20% (Kunkera et al. 2022). Given the strategic objective of Norwegian shipyards to reduce costs by 10-15% (Resyard 2023), the potential productivity gains from 3D modeling could contribute significantly towards achieving this target. Thus, the adoption of digital twin technology, specifically 3D modeling, could pave the way for operational advancements and cost efficiencies in the shipyard.

The incremental way of introducing and applying the digital twin concept into Norwegian shipyards would not need the high investment cost in the beginning but rather build on applications when possible. The benefits and challenges can also be analyzed before moving forward with more applications. As there are not many proven digital twin applications in shipyards, it's important to keep in mind that it's not certain that benefits and challenges will be the same as in other industries. Despite its potential, DT technology is still in the early stages, and there are significant challenges to overcome, including costs, the complexity of information, maintenance, lack of standards and regulations, and cybersecurity and communication issues (Botın-Sanabria et al. 2022). As a specific example of a challenge in the maritime industry, Wen-Hao et al. (2021) reported that with the impact of variable loads and the high salt spray environment of the ocean, performance degradation was rapid, and failures occurred frequently.

The proposed method of gradually building a digital twin by starting with a few key applications and then integrating and building upon them is not only theoretically sound, but it is also endorsed by industry professionals. In semi-structured interviews conducted with representatives from the digital twin department of a renowned consulting firm, they validated this approach. According to these representatives, beginning with critical applications such as resource management and predictive maintenance provides a solid foundation upon which a more comprehensive digital twin can be constructed. Their insights further emphasized that this incremental approach can be particularly beneficial in managing the high initial investment costs and complexities associated with deploying a digital twin. Thus, their industry perspective lends additional credence to this step-by-step methodology for introducing and expanding the use of digital twins in Norwegian shipyards.

It's important to have in mind that the digital twin application should be designed with scalability in mind. This includes choosing flexible and robust software and hardware solutions that can accommodate future growth. A modular design strategy that allows for the gradual addition of functionalities can also be beneficial for scalability.

7.2 Research question 2

RQ2: How can digital twin technology influence decision-making processes in managing labor hours within shipyard operations?

Digital twin technology, as substantiated through the conducted case study, presents significant potential to reshape decision-making processes within shipyard operations, specifically concerning labor hour management. A salient attribute of this technology lies in its ability to automate the prediction of labor hours for forthcoming projects. Such automation not only boosts operational efficiency via swift and precise estimations but also reallocates human resources to more complex and intellectually demanding tasks. By incorporating digital twin technology, shipyards could potentially enhance process efficiency and optimize their workforce utilization.

Furthermore, the predictive capability of this technology extends beyond merely being a time-saving convenience. It plays a vital role in cost management, especially given the context of high labor costs in Norway. Experts in the field also inform that reasons for yards being closed are not always for not getting contracts, but may also result from substantial cost overruns. A single project surpassing its estimated budget excessively could precipitate a yard's compulsory shutdown.

Production costs are primarily influenced by investments and labor expenses. Given that materials and ship equipment are obtained from international markets where conditions tend to be uniform, labor costs in production become a significant differentiating factor (Pires Jr et al. 2009). Estimates from Hagen and Erikstad (2014) say that labor cost contributes to one-third of the cost of a vessel. Precise labor hour forecasting might serve as a crucial determinant, distinguishing between the financial triumph and downfall of a project. Over estimations could lead to overpriced proposals, thereby losing contracts, whereas underestimations may result in cost overruns and, in the worst case, the closing of a yard. By harnessing the precision of digital twin technology for labor hour predictions, shipyards can achieve more control over project costs and improve their competitive standing.

Additionally, digital twin technology could expedite the decision-making process. In the competitive realm of shipbuilding, the ability to provide accurate cost estimates swiftly can often sway contract negotiations favorably. By generating accurate labor hour predictions promptly, digital twin technology furnishes decision-makers with crucial information, enabling informed and timely decisions. This responsiveness could provide a competitive edge, allowing shipyards to seize potential opportunities more promptly.

Beyond its predictive capabilities, digital twin technology also serves as a tool for explanatory modeling, further enhancing its value in managing labor hours and cost. The models developed can shed light on the influential factors affecting labor hours, providing key insights into the intricacies of shipyard operations. By understanding the patterns, correlations, and causative factors in labor hour allocation, decision-makers can pinpoint areas of inefficiencies or bottlenecks and devise targeted strategies for improvement. For example, in the case study, it was concluded that CGT was the most important factor for labor hours used on a vessel. At the same time, the significant impact of repetitiveness, especially on the engineering and management departments, was shown.

The explanatory aspect of these models can assist in determining which parameters, when adjusted, could result in significant labor hour reductions. For instance, understanding the impact of variables such as level of repetition, building strategies, powerplant index, or outfitting weights could guide strategic decisions around these factors to minimize labor hours. Consequently, this leads to cost reductions and improvements in overall operational efficiency. These insights not only inform immediate decisions but also help in crafting long-term strategic plans.

For the digital twin concept to be implemented into labor hour management, there are some requirements, such as:

- Data Availability and Quality
- Integration with Existing Systems
- Technical Expertise
- Willingness to invest
- Scalability

It's important to have them in mind when implementing the models from the case study into a shipyard. Given the difficulty of generalizing the findings to other shipyards, it is more appropriate at this stage to focus on the yard studied in the case study. Liland 2023 pointed to a sufficient amount of quality data as the main requirement for both cost estimation and lead time prediction when writing about the potential of machine learning in Norwegian shipyards. The results from the model developed in the case study done in this thesis have shown that the data is both available and of sufficient quality to be able to predict labor hours with high precision. There has also been a demonstrated willingness to invest, as shown by the resources allocated to carry out this master thesis. With a strategic commitment to invest, it is feasible to source the requisite technical expertise externally.

Integration of existing systems is for the case study solved as both Python code and formulas can be implemented directly into the system used today, Tableau. Scalability is important when having a bigger goal further along, a digital twin that covers the whole shipyard. As the models are built in the second-most in-demand programming language (HackerRank 2020), it's logical to think that it would be able to join the scaling of a digital twin.

The models exhibit a relatively high R^2 score, but a pertinent question remains: is this score sufficiently high for the shipyard's requirements? Specifically, for total hours, the model yields a Mean Absolute Error (MAE) of 25,449 hours. Considering an hourly wage of 500 Norwegian Krone, this translates to a substantial cost implication of approximately 12.7 million NOK. It is critical to recognize that this MAE represents deviations in both directions; thus, it encompasses instances where the budget may either be underestimated or exceeded. Furthermore, as historical data on price predictions for the shipyard could not be procured at this juncture, a direct comparison with manual predictions, which might be based on experience, is unattainable. Nevertheless, an R^2 score of 0.935 is indicative of the model's potential to positively impact the decision-making process in managing hours more efficiently and reducing human error.

The practical implications of the findings for the shipbuilding industry are manifold. By determining the significant variables influencing labor hours in boat production, companies can strategically optimize their processes to minimize labor-intensive tasks. This is particularly relevant in Norway, where labor costs constitute a significant portion of overall costs. Through better management of labor hours, the industry can mitigate these high costs, potentially leading to competitive pricing and more sustainable business practices.

Furthermore, the research can influence the way companies allocate their labor resources. With a reliable predictive model, resource allocation can become a more data-driven process, reducing the reliance on subjective judgment and potentially leading to increased operational efficiency.

Moreover, the research opens the door for increased transparency in labor usage within the shipbuilding industry. With the advent of predictive models, stakeholders can better understand where labor hours are spent during the manufacturing process. This insight can be invaluable for customers as it provides a clearer picture of the production process, contributing to more informed purchasing decisions.

Digital twin technology has a transformative potential to reshape how labor hours are managed within shipyard operations. Beyond the accurate prediction of labor hours, this technology has implications for the distribution and allocation of human resources. For example, the improved accuracy in labor hour predictions could enable more efficient workforce scheduling. With the precision afforded by digital twins, the allocation of workers could be fine-tuned to better align with project needs, reducing idle time and boosting productivity. Furthermore, this precision could be leveraged to improve work-life balance for the employees, by reducing instances of overwork or underwork. Thus, digital twin technology does not only optimize operations but also contributes to creating a more equitable and sustainable work environment.

8 Conclusion

This chapter summarises the key findings of the master thesis and presents concluding reflections. Additionally, it reflects on the limitations encountered during the research process, and lastly, suggestions for further research are presented.

The primary objective of this master thesis is to investigate how to introduce and implementing digital twin technology in Norwegian shipyards, with a special focus on assessing its potential in operational improvement and analyzing its impact on decision-making processes, particularly in the management of labor hours. To achieve this, the following research questions were posed: RQ1: How can the concept of a digital twin be introduced and applied in Norwegian shipyards to improve their operations? and RQ2: How can digital twin technology influence decision-making processes in managing labor hours within shipyard operations? Through this inquiry, the thesis has generated structured insights into operational management within Norwegian shipyards, and proposed a framework for implementing digital twin technology to enhance operational efficiencies and decision-making capabilities.

Given the global competitive pressures, especially from regions with lower labor costs, and the influence of market factors such as oil price fluctuations, it is imperative for Norwegian shipyards to achieve cost-effectiveness (Hagen and Erikstad 2014. These yards operate within a complex Engineer-to-Order (ETO) environment, characterized by low levels of digitalization and automation, necessitating efficient cost management strategies for survival and growth (Strandhagen 2022).

Norwegian yards can still be globally competitive in the most advanced segments if they can reduce costs by 10-15 percent (Resyard 2023). The adoption of digital twin technology, as demonstrated in this thesis, can catalyze improvements in performance metrics, including production cost, building time, and quality. However, given the myriad challenges, such as the intricate ETO environment and limited budgets, the thesis advocates an incremental approach. This entails initializing the digital twin adoption with a few critical applications and subsequently expanding and integrating based on the insights gained. In the current state of Norwegian shipyards, resource management emerges as the paramount domain for deployment, exhibiting substantial promise in terms of applicability and potential impact.

Furthermore, digital twins' capacity to offer virtual replicas of physical systems enhances real-time surveillance and data analytics, consequently bolstering decision-making (Wang et al. 2022). The thesis identifies resource management, predictive maintenance, Disaster and emergency response, and crane and equipment utilization as potential domains for digital twin integration. Digital twins' efficacy is amplified when various applications are combined to create a complete virtual model of the shipyard. For example, resource management and predictive maintenance have their own distinct advantages. However, integrating these applications can lead to even greater benefits as it allows for advanced simulations that consider both the availability of labor hours and the utilization of equipment.

As digital twin technology is still evolving, it is critical for shipyards to build scalable solutions and be vigilant regarding the challenges like costs, complexity, maintenance, and cybersecurity. The thesis recommends a modular and scalable approach that is endorsed by industry expert, with continuous evaluations of the rewards and difficulties as the technology is progressively deployed.

Production costs are primarily influenced by investments and labor expenses. Given that materials and ship equipment are obtained from international markets where conditions tend to be uniform, labor costs in production become a significant differentiating factor(Pires Jr et al. 2009). This Master's thesis has, therefore, also explored the impact of digital twin technology on decision-making processes related to the management of labor hours within shipyard operations. This was done partly through a case study was done by using digital twin technology to estimate labor hours needed in different departments for new potential projects.

The case study found the available data of enough quality and predicted total hours for new vessels with a R^2 value of 0.935 on unseen data. The application of this technology demonstrated its potential to transform labor hour management within shipyard operations. It introduced automation to the prediction process, leading to more efficient and accurate estimations that enable better resource allocation. The digital twin's predictive capabilities also offer substantial benefits in cost management, a particularly important aspect given the high labor costs in Norwegian shipyards. Moreover, the technology's capacity for explanatory modeling yielded valuable insights into factors affecting labor hours, paving the way for more strategic decision-making to optimize processes and reduce costs.

However, for successful implementation, requirements such as data availability and quality, integration with existing systems, technical expertise, willingness to invest, and solubility must be met. Hence, the digital twin technology presents an innovative approach that can potentially enhance efficiency, cost management, and overall decision-making within shipyard operations, given the prerequisites are adequately addressed.

This study has effectively addressed the research questions posed, providing a robust understanding of the potential and implications of digital twin technology within the Norwegian shipyard industry. It discusses potential areas of digital twin that can be integrated into shipyards. The conclusion is based on a literature study, interviews, and a case study. It suggests that Norwegian shipyards should start with resource management and then gradually adding new applications, using an incremental approach.

8.1 Limitations

Due to limited time and wide research scope, the research was performed with a low level of detail, and many parameters were not looked into. The wide scope also compromised the depth of the study.

The case study was performed in a specific Norwegian shipyard, and the results might not necessarily apply to all shipyards globally. This is due to varying operational practices, labor costs, and regulatory frameworks.

The current research does not delve deep into potential ethical and legal considerations, such as data privacy, intellectual property rights, or potential job displacement due to automation.

Interviews constructed the methodology. This has some limitations:

- Bias: Interviewers may have biases that influence their interpretation of the responses they receive, and their own biases may be reflected in the questions they ask.
- Reliability: Interviews may not be reliable because the same information may be interpreted differently by different interviewers or reported differently by other interviewees. Interviews can always be questioned due to the involvement of humans' perception and comprehension

of reality.

- Lack of control: The researcher has limited control over the responses they receive during an interview, as they depend on the interviewee's willingness to share information.
- Validity: It can be challenging to determine the validity of the information gathered through interviews, as the responses may be influenced by the interviewer or the interviewee's desire to present themselves in a certain way.

The interviews were supplemented with a literature study that helped limit these limitations. However, some details were not possible to obtain through the literature study.

Only one shipyard was part of the case study and interviews; therefore hard to generalize the research to all Norwegian shipyards.

8.2 Further research

Further data collection is needed to refine and improve the predictive models. It should also include collaboration with several shipyards to be able to generalize the results.

Further research should contain more practical implementations of applications of digital twin that are proven in other similar industries. It should also contain research around merging applications together to test out the incremental approach that is suggested in this master thesis. As digital twin technology gets integrated into more areas of shipyard operations, understanding how to scale these solutions becomes crucial. Future research can explore best practices for planning and managing the growth of digital twin implementations.

Further research should expand the case study to other areas of shipyard resource management to include physical resources as well.

Bibliography

Aivaliotis, Panagiotis, Konstantinos Georgoulias and George Chryssolouris (2019). 'The use of Digital Twin for predictive maintenance in manufacturing'. In: *International Journal of Computer Integrated Manufacturing* 32.11, pp. 1067–1080.

Alexopoulos, Evangelos C (2010). 'Introduction to multivariate regression analysis'. In: *Hippokratia* 14.Suppl 1, p. 23.

Alhaidari, Fahd et al. (2021). 'Intelligent software-defined network for cognitive routing optimization using deep extreme learning machine approach'. In.

Almalki, Sami (2016). 'Integrating Quantitative and Qualitative Data in Mixed Methods Research–Challenges and Benefits.' In: *Journal of education and learning* 5.3, pp. 288–296.

Amdam, Rolv Petter and Ove Bjarnar (2015). 'Globalization and the development of industrial clusters: Comparing two Norwegian clusters, 1900–2010'. In: *Business History Review* 89.4, pp. 693–716.

Amiri, Roohollah et al. (2018). 'A machine learning approach for power allocation in HetNets considering QoS'. In: *2018 IEEE international conference on communications (ICC)*. IEEE, pp. 1–7.

Andritsos, Fivos and Juan Perez-Prat (2000). 'The automation and integration of production processes in shipbuilding'. In: *State-of-the-Art report, Joint Research Centre. European Commission, Europe.*

Arthur, David and Sergei Vassilvitskii (2007). 'K-means++: The advantages of careful seeding'. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, pp. 1027–1035.

Ashton, Kevin et al. (2009). 'That 'internet of things' thing'. In: *RFID journal* 22.7, pp. 97–114.

Association, World Medical (2013). 'WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects'. In.

Bastani, Hamsa, Dennis J Zhang and Heng Zhang (2022). 'Applied machine learning in operations management'. In: *Innovative Technology at the Interface of Finance and Operations*. Springer, pp. 189–222.

Baxter, Pamela and Susan Jack (2008). 'Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers'. In: *The Qualitative Report* 13.4, pp. 544–559.

Becker, Saul, Alan Bryman and Harry Ferguson (2012). *Understanding research for social policy and social work: themes, methods and approaches*. policy press.

Bicheno, John and Matthias Holweg (2016). *The lean toolbox: The essential guide to lean transformation*. PICSIE Books.

Billinghurst, Mark and Andreas Duenser (2018). 'Augmented reality in education'. In: *Cambridge Handbook of Learning Sciences*. Cambridge University Press, pp. 643–656.

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. Springer.

Botın-Sanabria, Diego M et al. (2022). 'Digital twin technology challenges and applications: A comprehensive review'. In: *Remote Sensing* 14.6, p. 1335.

Bourne, Mike CS et al. (2018). 'Designing, implementing and updating performance measurement systems'. In: *International Journal of Operations & Production Management* 20.7, pp. 754–771.

Bowen, Glenn A. (2009). 'Document Analysis as a Qualitative Research Method'. In: *Qualitative Research Journal* 9.2, pp. 27–40.

Brahme, Anders (2014). *Comprehensive biomedical physics*. Newnes.

Breiman, Leo (2001). 'Random forests'. In: *Machine learning* 45.1, pp. 5–32.

Bryman, Alan (2012). *Social research methods*. 4th ed. Oxford University Press.

Brynjolfsson, Erik and Andrew McAfee (2017). 'The Business of Artificial Intelligence: What it can — and cannot — do for your organization'. In: *Harvard Business Review*.

Bzdok, Danilo, Martin Krzywinski and Naomi Altman (2017). 'Machine learning: a primer'. In: *Nature methods* 14.12, p. 1119.

Campbell, Donald T and Julian C Stanley (1963). *Experimental and quasi-experimental designs for research*. Wadsworth Publishing.

Cannas, Violetta Giada and Jonathan Gosling (2021). 'A decade of engineering-to-order (2010–2020): Progress and emerging themes'. In: *International Journal of Production Economics* 241, p. 108274.

Casino, Fran, Thomas K Dasaklis and Constantinos Patsakis (2019). 'A systematic literature review of blockchain-based applications: Current status, classification and open issues'. In: *Telematics and Informatics* 36, pp. 55–81.

Chai, Tzu-Yi and Jen-Cheng Lin (2021). 'A comparison of linear and non-linear regression techniques in modeling water quality parameters'. In: *Water* 13.8, p. 1103.

Chapman, Pete et al. (2000). 'CRISP-DM 1.0: Step-by-step data mining guide'. In: *SPSS inc* 9.13, pp. 1–73.

Chen, Injazz J (2001). 'Planning for ERP systems: analysis and future trend'. In: *Business process management journal*.

Creswell, John W. and Cheryl N. Poth (2018). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. 4th ed. SAGE Publications.

Croom, Simon (2010). 'Introduction to research methodology in operations management'. In: *Researching operations management*. Routledge, pp. 56–97.

Cutler, David R et al. (2007). 'Random forests for classification in ecology'. In: *Ecology* 88.11, pp. 2783–2792.

Dai, Bin et al. (2021). 'Routing optimization meets Machine Intelligence: A perspective for the future network'. In: *Neurocomputing* 459, pp. 44–58.

Dayhoff, Judith E and James M DeLeo (2001). 'Artificial neural networks: opening the black box'. In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 91.S8, pp. 1615–1635.

Deng, Li, Dong Yu et al. (2014). 'Deep learning: methods and applications'. In: *Foundations and trends® in signal processing* 7.3–4, pp. 197–387.

Denzin, Norman K (1978). 'The research act: A theoretical introduction to sociological methods'. In.

developers, Scikit-learn (2021). *Cross-validation: evaluating estimator performance*. Accessed: 2023-05-22. URL: https://scikit-learn.org/stable/modules/cross_validation.html.

DiCicco-Bloom, Barbara and Benjamin F. Crabtree (2006). 'The Qualitative Research Interview'. In: *Medical Education* 40.4, pp. 314–321.

Dou, Zixin et al. (2021). 'Regional Manufacturing Industry Demand Forecasting: A Deep Learning Approach'. In: *Applied Sciences* 11.13, p. 6199.

Draper, NR and H Smith (1998). *Applied regression analysis*. Wiley-Interscience.

Dugnas, Karolis and Oddmund Oterhals (2008). 'State-of-the-art shipbuilding: towards unique and integrated lean production systems'. In: *Proceedings of IGLC16: 16th annual conference of the international group for lean construction*, pp. 321–331.

El Naqa, Issam and Martin J Murphy (2015). 'What is machine learning?' In: *machine learning in radiation oncology*. Springer, pp. 3–11.

Everitt, Brian S and Anders Skrondal (2010). 'The Cambridge dictionary of statistics'. In.

Farrar, Donald E and Robert R Glauber (1967). 'Multicollinearity in Regression Analysis: The Problem Revisited'. In: *The Review of Economics and Statistics* 49.1, pp. 92–107.

Field, Andy (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Fleckenstein, Mike, Lorraine Fellows and Krista Ferrante (2018). *Modern data strategy*. Springer.

Flyvbjerg, Bent (2006). 'Five Misunderstandings About Case-Study Research'. In: *Qualitative Inquiry* 12.2, pp. 219–245.

Frank, Alejandro Germán, Lucas Santos Dalenogare and Néstor Fabián Ayala (2019). 'Industry 4.0 technologies: Implementation patterns in manufacturing companies'. In: *International Journal of Production Economics* 210, pp. 15–26.

Al-Fuqaha, Ala et al. (2015). 'Internet of things: A survey on enabling technologies, protocols, and applications'. In: *IEEE communications surveys & tutorials* 17.4, pp. 2347–2376.

Fürber, Christian (2016). 'Semantic Technologies'. In: *Data Quality Management with Semantic Technologies*. Springer, pp. 56–68.

Gacek, Stanisław (2018). 'Due date assignment using neural networks for standard products in small batch and multi assortment make-to-order company'. In: *Proceedings of the Carpathian Logistics Congress, Prague, Czech Republic*, pp. 3–5.

Géron, Aurélien (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc."

Gerring, John (2006). *Case Study Research: Principles and Practices*. Cambridge University Press.

Glaessgen, Edward and David Stargel (2012). 'The digital twin paradigm for future NASA and US Air Force vehicles'. In: *53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA*, p. 1818.

Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). *Deep Learning*. MIT Press.

Goodman, Leo A (1961). 'Snowball sampling'. In: *The annals of mathematical statistics*, pp. 148–170.

Grieves, Michael and John Vickers (2014). 'Digital twin: Manufacturing excellence through virtual factory replication'. In: *White paper* 1.201, pp. 1–7.

— (2017). 'Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems'. In: *Transdisciplinary perspectives on complex systems: New findings and approaches*, pp. 85–113.

Groover, Mikell P. (2008). *Automation, Production Systems, and Computer-Integrated Manufacturing*. New Jersey: Prentice Hall Press.

Guyon, Isabelle and André Elisseeff (2003). 'An Introduction to Variable and Feature Selection'. In: *Journal of Machine Learning Research* 3, pp. 1157–1182.

HackerRank (2020). *2020 Developer Skills Report*. Technical Report. HackerRank. URL: https://info.hackerrank.com/rs/487-WAY-049/images/HackerRank-2020-Developer-Skills-Report.pdf.

Hagen, Arnulf and Stein Ove Erikstad (2014). 'Shipbuilding'. In: *Trondheim: Norwegian University of Science and Technology*.

Hair, Joseph F (2009). 'Multivariate data analysis'. In.

Han, Jiawei, Micheline Kamber and Jian Pei (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hawkins, Douglas M (2004). 'The problem of overfitting'. In: *Journal of chemical information and computer sciences* 44.1, pp. 1–12.

Hedmond, Shane (Aug. 2020). *OpenSpace launches AI-powered production tracking and object search from 360 degree photos*. URL: https://www.constructionjunkie.com/blog/2020/8/5/openspace-launches-ai-powered-production-tracking-and-object-search-from-360-degree-photos.

Hellström, Thomas, Virginia Dignum and Suna Bensch (2020). 'Bias in Machine Learning–What is it Good for?' In: *arXiv preprint arXiv:2004.00686*.

Hinton, Geoffrey and Terrence J Sejnowski (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

Hjartholm, Bendik Thune (2019). 'Augmented Reality for Operator Support in Norwegian Shipyards: A Study of Applications, Benefits and Challenges'. MA thesis. NTNU.

Inc., TIBCO Software (n.d.).

Jain, Anil K, M Narasimha Murty and Patrick J Flynn (2010). 'Data clustering: 50 years beyond K-means'. In: *Pattern recognition letters* 31.8, pp. 651–666.

Kaelbling, Leslie Pack, Michael L Littman and Andrew W Moore (1996). 'Reinforcement learning: A survey'. In: *Journal of artificial intelligence research* 4, pp. 237–285.

Kaluzny, Bohdan L et al. (2011). 'An application of data mining algorithms for shipbuilding cost estimation'. In: *Journal of Cost Analysis and Parametrics* 4.1, pp. 2–30.

Karlsson, Christer (2010). 'Researching operations management'. In: *Researching operations management*. Routledge, pp. 20–55.

Kawulich, Barbara B. (2005). 'Participant Observation as a Data Collection Method'. In: *Forum: Qualitative Social Research* 6.2.

Kerzner, Harold (2017). *Project management: A systems approach to planning, scheduling, and controlling*. John Wiley & Sons.

Kozjek, Dominik, Borut Rihtaršič, Peter Butala et al. (2018). 'Big data analytics for operations management in engineer-to-order manufacturing'. In: *Procedia CIRP* 72, pp. 209–214.

Kritzinger, Werner et al. (2018). 'Digital Twin in manufacturing: A categorical literature review and classification'. In: *Ifac-PapersOnline* 51.11, pp. 1016–1022.

Kunkera, Zoran et al. (2022). 'Using Digital Twin in a Shipbuilding Project'. In: *Applied Sciences* 12.24, p. 12721.

Labs, Sharp Sight (2018). *How to use 'train$_t$est$_s$plit' $fromscikit - learn$*. Accessed: 2023-04-27. URL: https://www.sharpsightlabs.com/blog/scikit-train_test_split/.

Lamb, Thomas (2004). *Ship Design and Construction, Volumes 1-2*. Society of Naval Architects and Marine Engineers (SNAME).

Lamb, Thomas and Aasmund Hellesoy (2002). 'A shipbuilding productivity predictor'. In: *Journal of ship production* 18.02, pp. 79–85.

Lee, J. et al. (2013). 'Recent advances and trends in predictive manufacturing systems in big data environment'. In: *Manufacturing Letters* 1.1, pp. 38–41.

Leng, Jiewu et al. (2020). ‘Digital twin-driven rapid reconfiguration of the automated manufacturing system via an open architecture model’. In: *Robotics and Computer-Integrated Manufacturing* 63, p. 101895.

Lever, Jake, Martin Krzywinski and Naomi Altman (2016). ‘Points of significance: model selection and overfitting’. In: *Nature methods* 13.9, pp. 703–705.

Liaw, Andy and Matthew Wiener (2002). ‘Classification and regression by randomForest’. In: *R news* 2.3, pp. 18–22.

Liker, Jeffrey K and Thomas Lamb (2000). *Lean Manufacturing Principles Guide, Version 0.5. A Guide to Lean Shipbuilding.* Tech. rep. MICHIGAN UNIV ANN ARBOR.

Liland, E. (2023). ‘The Potential of Machine Learning in Norwegian Shipyard Operations’. NTNU, Trondheim, Norway.

Lim, Kendrik Yan Hong, Pai Zheng and Chun-Hsien Chen (2020). ‘A state-of-the-art survey of Digital Twin: techniques, engineering product lifecycle management and business innovation perspectives’. In: *Journal of Intelligent Manufacturing* 31, pp. 1313–1337.

Lingitz, Lukas et al. (2018). ‘Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer’. In: *Procedia Cirp* 72, pp. 1051–1056.

Lu, Yinhai et al. (2017). ‘Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues’. In: *Robotics and Computer-Integrated Manufacturing* 49, pp. 1–14.

MacQueen, James B (1967a). ‘Some Methods for classification and Analysis of Multivariate Observations’. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1.14, pp. 281–297.

— (1967b). ‘Some methods for classification and analysis of multivariate observations’. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14, pp. 281–297.

Mahanti, Rupa (2019). *Data quality: dimensions, measurement, strategy, management, and governance*. Quality Press.

Maneejuk, Paravee and Woraphon Yamaka (2021). ‘Significance test for linear regression: how to test without P-values?’ In: *Journal of Applied Statistics* 48.5, pp. 827–845.

Manikas, Andrew et al. (2020). ‘A review of operations management literature: a data-driven approach’. In: *International Journal of Production Research* 58.5, pp. 1442–1461.

Mannino, Michael, Yanjuan Yang and Young Ryu (2009). ‘Classification algorithm sensitivity to training data with non representative attribute noise’. In: *Decision Support Systems* 46.3, pp. 743–751.

Matthews, Robert and Elizabeth Ross (2010). *Research methods: A practical guide for the social sciences*. Pearson Education Ltd.

McKinney, Wes (2010). ‘Data structures for statistical computing in python’. In: *In Proceedings of the 9th Python in Science Conference* 445, pp. 51–56.

Mellbye, CS, Anders M Helseth and Erik W Jakobsen (2018). ‘Maritim verdiskapingsbok 2018’. In: *Menon Economics*.

Mezzogori, Davide, Giovanni Romagnoli and Francesco Zammori (2019). ‘Deep learning and WLC: how to set realistic delivery dates in high variety manufacturing systems’. In: *IFAC-PapersOnLine* 52.13, pp. 2092–2097.

Miles, Jeremy (2005). ‘R-squared, adjusted R-squared’. In: *Encyclopedia of statistics in behavioral science*.

Monostori, László, Ben van der Zee and Ove Isaksson (2016). 'Cyber-physical systems in manufacturing'. In: *CIRP Annals* 65.2, pp. 621–641.

Montgomery, Douglas C, Elizabeth A Peck and G Geoffrey Vining (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Morris, H et al. (2009). 'The Association of Business Schools: Academic journal quality guide, version 3'. In.

Moyst, Howard and Biman Das (2005). 'Factors affecting ship design and construction lead time and cost'. In: *Journal of ship production* 21.03, pp. 186–194.

Muhammad, Iqbal and Zhu Yan (2015). 'SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY.' In: *ICTACT Journal on Soft Computing* 5.3.

Muller, Ralf (2017). *Project governance*. Taylor & Francis.

Murphy, Rory et al. (2019). 'Machine learning technologies for order flowtime estimation in manufacturing systems'. In: *Procedia CIRP* 81, pp. 701–706.

Nautiyal, D (2019). *Underfitting and Overfitting in Machine Learning*. GeeksforGeeks. Accessed December 14, 2018. URL: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/.

Negri, Elisa, Marco Fumagalli and Marco Macchi (2017). 'A review of the roles of digital twin in CPS-based production systems'. In: *Procedia Manufacturing* 11, pp. 939–948.

Neter, John et al. (1996). 'Linear Regression'. In: *Applied Linear Statistical Models* 4, pp. 0–0.

Noura, Mahda, Mohammed Atiquzzaman and Martin Gaedke (2019). 'Interoperability in internet of things: Taxonomies and open challenges'. In: *Mobile networks and applications* 24, pp. 796–809.

Olhager, Jan (2003). 'Strategic positioning of the order penetration point'. In: *International journal of production economics* 85.3, pp. 319–329.

Onaji, Igiri et al. (2022). 'Digital twin in manufacturing: conceptual framework and case studies'. In: *International journal of computer integrated manufacturing* 35.8, pp. 831–858.

Oterhals, Oddmund, Gøran Johannessen and Arild Hervik (2011). *STX OSV. Ringvirkninger av verftsvirksomheten i Norge*.

Pedregosa, Fabian et al. (2011). 'Scikit-learn: Machine learning in Python'. In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830.

Pfeiffer, András et al. (2016). 'Manufacturing lead time estimation with the combination of simulation and statistical learning methods'. In: *Procedia CIRP* 41, pp. 75–80.

Pires, Flávia et al. (2019). 'Digital twin in industry 4.0: Technologies, applications and challenges'. In: *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*. Vol. 1. IEEE, pp. 721–726.

Pires Jr, Floriano, Thomas Lamb and Cassiano Souza (2009). 'Shipbuilding performance benchmarking'. In: *International journal of business performance management* 11.3, pp. 216–235.

Premsankar, Gopika, Mario Di Francesco and Tarik Taleb (2018). 'Edge computing for the Internet of Things: A case study'. In: *IEEE Internet of Things Journal* 5.2, pp. 1275–1284.

Qi, Qinglin and Fei Tao (2018). 'Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison'. In: *Ieee Access* 6, pp. 3585–3593.

Qinghua, Gao and Wang Chunjiang (2019). 'Digital twin in shipbuilding'. In: *China Shipbuilding* 1, pp. 1–10.

Rajasekar, S, P Philominathan and V Chinnathambi (2006). *Research methodology, Ar XIV Physics*.

Resyard (2023). 'Project proposal submitted 15.2.2023 to the Research Council of Norway'. NTNU, Trondheim, Norway.

Robins, Mark (2017). 'The Future of Deep Learning: Challenges & Solutions'. In: *Proceedings of the Computing Frontiers Conference*, pp. ii–ii.

Rosen, Roland et al. (2015). 'About the importance of autonomy and digital twins for the future of manufacturing'. In: *Ifac-Papersonline* 48.3, pp. 567–572.

Russell, Stuart and Peter Norvig (2020). *Artificial Intelligence: A Modern Approach.* 4th ed. Pearson.

Schmidhuber, Jürgen (2015). 'Deep learning in neural networks: An overview'. In: *Neural networks* 61, pp. 85–117.

Schneckenreither, Manuel, Stefan Haeussler and Christoph Gerhold (2021). 'Order release planning with predictive lead times: a machine learning approach'. In: *International Journal of Production Research* 59.11, pp. 3285–3303.

Semini, Marco, Per Olaf Brett, Arnulf Hagen et al. (2018). 'Offshoring strategies in Norwegian ship production'. In: *Journal of Ship Production and Design* 34.01, pp. 59–71.

Semini, Marco, Per Olaf Brett, Jan Ola Strandhagen et al. (2022). 'Comparing Offshore Support Vessel Production Times between Different Offshoring Strategies Practiced at Norwegian Shipyards'. In: *Journal of Ship Production and Design* 38.02, pp. 76–88.

Semini, Marco, Dag E Gotteberg Haartveit et al. (2014). 'Strategies for customized shipbuilding with different customer order decoupling points'. In: *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment* 228.4, pp. 362–372.

Shi, Weisong et al. (2016). 'Edge computing: Vision and challenges'. In: *IEEE Internet of Things Journal* 3.5, pp. 637–646.

Shorten, Connor and Taghi M Khoshgoftaar (2019). 'A survey on image data augmentation for deep learning'. In: *Journal of big data* 6.1, pp. 1–48.

Shuttleworth, Martyn (2008). *Case Study Research Design.* https://explorable.com/case-study-research-design. Accessed: 2023-03-31.

Si, Jinpeng et al. (2019). 'Research on the application of digital twin technology in shipbuilding engineering'. In: *Journal of Marine Science and Engineering* 7.9, p. 289.

Slack, Nigel, Alistair Brandon-Jones and Robert Johnston (2018). *Operations management.* Pearson UK.

Stankovic, John A (2014). 'Research directions for the internet of things'. In: *IEEE internet of things journal* 1.1, pp. 3–9.

Statista (Nov. 2022). *Active shipyards worldwide 2014-2022.* URL: https://www.statista.com/statistics/1102673/active-shipyards-worldwide/.

Steck, Harald (2019). 'Embarrassingly shallow autoencoders for sparse data'. In: *The World Wide Web Conference*, pp. 3251–3257.

Stensvold, T. (2022). 'Derfor Er Det Krise for Norsk Skipsbygging'. In: *Teknisk ukeblad* 01.22.

Stephansen-Smith, Finn Julius (2020). 'Using Neural Networks for IoT Power Management'. MA thesis. NTNU.

Stevenson, William J, Mehran Hojati and James Cao (2014). *Operations management.* McGraw-Hill Education New York.

Strandhagen (2022). 'Towards Next-Generation Yard Logistics'. In.

Strandhagen, Jo Wessel et al. (2019). 'Digitalized manufacturing logistics in engineer-to-order operations'. In: *IFIP International Conference on Advances in Production Management Systems*. Springer, pp. 579–587.

Strobl, Carolin et al. (2007). 'Bias in random forest variable importance measures: Illustrations, sources and a solution'. In: *BMC Bioinformatics* 8.1, p. 25.

Su, S Y and D Y Sha (2004). 'Due date assignment using artificial neural networks under different shop floor control strategies'. In: *International Journal of Production Research* 42.9, pp. 1727–1745.

Süve, Mustafa Furkan, Cengiz Gezer and Gökhan İnce (2022). 'Predictive maintenance framework for production environments using digital twin'. In: *Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, held August 24-26, 2021. Volume 2*. Springer, pp. 455–462.

Taleb, Ikbal, Mohamed Adel Serhani and Rachida Dssouli (2018). 'Big data quality: A survey'. In: *2018 IEEE International Congress on Big Data (BigData Congress)*. IEEE, pp. 166–173.

Tao, Fei et al. (2018). 'Digital twin-driven product design, manufacturing and service with big data'. In: *The International Journal of Advanced Manufacturing Technology* 94, pp. 3563–3576.

Uhlemann, Thomas J.H., Egon Müller and Matthias Meyer (2017). 'The digital twin: Realizing the cyber-physical production system for industry 4.0'. In: *Procedia CIRP* 61, pp. 335–340.

Van Der Walt, Stéfan, S Chris Colbert and Gaël Varoquaux (2011). 'The NumPy array: a structure for efficient numerical computation'. In: *Computing in Science & Engineering* 13.2, pp. 22–30.

Varghese, Alex and Bala Tamma (2018). 'Digital twin: Values, challenges and enablers from a modeling perspective'. In: *Systems* 6.4, p. 42.

Verhagen, Mark D (2021). 'Identifying Model Complexity: A Machine Learning Framework'. In.

Virtanen, Pauli et al. (2020). 'SciPy 1.0: fundamental algorithms for scientific computing in Python'. In: *Nature methods* 17.3, pp. 261–272.

Voss, Christopher A, Nikos Tsikriktsis and Mark Frohlich (2002). 'Case research in operations management'. In: *International Journal of Operations & Production Management* 22.2, pp. 195–219.

Wang, Kan, Qianqian Hu and Jialin Liu (2022). 'Digital twin-driven approach for process management and traceability towards ship industry'. In: *Processes* 10.6, p. 1083.

Wen-Hao, WU, Chen Guo-bing and Yang Zi-chun (2021). 'The application and challenge of digital twin technology in ship equipment'. In: *Journal of Physics: Conference Series*. Vol. 1939. 1. IOP Publishing, p. 012068.

West, Jeremy, Dan Ventura and Sean Warnick (2007). 'Spring research presentation: A theoretical foundation for inductive transfer'. In: *Brigham Young University, College of Physical and Mathematical Sciences* 1.08.

Whitmore, Andrew, Anurag Agarwal and Li Da Xu (2015). 'The Internet of Things—A survey of topics and trends'. In: *Information systems frontiers* 17, pp. 261–274.

Winkelhaus, Sven and Eric H Grosse (2020). 'Logistics 4.0: a systematic review towards a new logistics system'. In: *International Journal of Production Research* 58.1, pp. 18–43.

Wu, Shaomin (2013). 'A review on coarse warranty data and analysis'. In: *Reliability Engineering & System Safety* 114, pp. 1–11.

Yang, Chen, Weiming Shen and Xianbin Wang (2018). 'The internet of things in manufacturing: Key issues and potential applications'. In: *IEEE Systems, Man, and Cybernetics Magazine* 4.1, pp. 6–15.

Yin, Robert K. (2014). ‘Case Study Research: Design and Methods’. In.

Zennaro, Ilenia et al. (2019). ‘Big size highly customised product manufacturing systems: a literature review and future research agenda’. In: *International Journal of Production Research* 57.15-16, pp. 5362–5385.

Zhangpeng, Gao and Matthew Flynn (2006). ‘Productive shipyards’. In: *Lloyd’s Shipping Economist* 28.6.

Zheng, Ting et al. (2021). ‘The applications of Industry 4.0 technologies in manufacturing context: a systematic literature review’. In: *International Journal of Production Research* 59.6, pp. 1922–1954.

Zhihan, Haibin and Mikael Fridenfalk (2023). ‘Digital Twins in the Marine Industry’. In: *Electronics* 12.9, p. 2025.

Appendix

A Apendix A: interview guide

Introduction

1. Thank and welcome the participant and ask if the interview can be recorded.
2. Ask if the participant name and professional information be disclosed in the project paper.
3. Ask interviewee about their experience in the Norwegian shipbuilding industry.
4. Ask whether it is possible for the participant to verify the accuracy of the information transcribed.
5. provide an explanation of the research topic, including the operational definitions that will be used. Additionally, please clarify the aim and scope of the study.

Main Part

1. Can you go through the main administrative operations and how they are conducted from the first customer encounter to delivery? Give examples if the participant doesn't know how to answer this question:

- Lead time prediction
- cost estimation
- Process estimation and tracking
- Resource estimation
- Quality control
- risk estimation
- Demand forecast
- ETC.

2. What do you think are the main challenges for the introduction of the digital twin concept in Norwegian shipyards? How are these challenges different from other industries?

3. Which industries do you find similar to the Norwegian shipbuilding industry?

4. How much data is available at a Norwegian shipyard, and where is most data gathered?

5. Why do you think Shipyards are far behind other industries from a digitalization point of view?

6. What areas do you find the most potential for digital twin applications at Norwegian shipyards?

7. Go through some digital twin applications and ask their opinions of the fit into operational management in Norwegian shipyards.

Closing

1. Ask if participants have anything to add to the topic discussed.
2. Ask if further communication is possible if more information is needed.
3. Thank the participant for the time.

B Apendix: Code - Correlation with all factors and departments

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the data
df = pd.read_csv('Historic_vessel_datat.csv')

# List of factors to check
factors = [
    "SERIES", "GT", "Compactness", "Engine_KW", "Engine_Nos", "Engine_Type",
    "No_thrusters", "Settlweight", "AccomodationA", "B_Strategy",
    "Crowdness", "Outfitweigh", "Cabins", "DP_Ind", "CGT",
    "POB", "Density_i", "Powerplant_i2",
    "Powerplant_i", "Accom_i", "Segmentex_i",
    "Experience_i2", "Density_i2", "Market_i"
]

# Target values
targets = ["Total_hours", "Engineering", "P&M",
           "Outfitting", "Scaffolding", "Project_management", "Outfitting_yard"]

# Calculate correlation for each target and print
correlation_dict = {}
for target in targets:
    print(f"\nCorrelation with {target}:")
    correlations = df[factors + [target]].corr()[target]

    # Exclude the target-target correlation
    correlations = correlations[correlations.index != target]

    print(correlations)

    correlation_dict[target] = correlations

# Convert the dictionary to a DataFrame for visualization
correlation_df = pd.DataFrame(correlation_dict)

# Calculate the average for each row
correlation_df['Average'] = correlation_df.mean(axis=1)

print("\nAverage correlations:")
print(correlation_df['Average'])
```

```
# Create a figure and a set of subplots
fig, ax = plt.subplots(figsize=(10, 8))

# Create a heatmap
sns.heatmap(correlation_df, annot=True, fmt=".2f", linewidths=.5, ax=ax, cmap='coolwarm')

# Show the plot
plt.show()
```

C Apendix: linear regression 1

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_absolute_error, make_scorer
from sklearn.model_selection import KFold
import matplotlib.pyplot as plt

# Read the CSV file
data = pd.read_csv("Historic_vessel_data.csv")

# Define the feature and target columns
feature_col = "CGT"
target_cols = [
    "Total_hours",
    "Outfitting_yard",
    "Project_management",
    "Production_scaffolding",
    "Production_outfitting",
    "Piping_&_machinery",
    "Engineering",
]

# Initialize Linear Regression
regressor = LinearRegression()

# Initialize KFold cross-validation
kf = KFold(n_splits=5, random_state=42, shuffle=True)

for target_col in target_cols:
    # Extract the feature and target
    X = data[feature_col].values.reshape(-1, 1)
    y = data[target_col].values.reshape(-1, 1)

    r2_scores = []
    mae_scores = []
    plt.figure()
    plt.scatter(X, y, color="blue", label="Data")

    for i, (train_index, test_index) in enumerate(kf.split(X)):
        X_train, X_test = X[train_index], X[test_index]
        y_train, y_test = y[train_index], y[test_index]
```

```
        regressor.fit(X_train, y_train)
        y_pred = regressor.predict(X_test)

        r2_scores.append(r2_score(y_test, y_pred))
        mae_scores.append(mean_absolute_error(y_test, y_pred))

        plt.plot(X_test, y_pred, color=plt.cm.viridis(i / 4.), label=f"Fold {i + 1}")

    print(f"Average R^2 score for {target_col}: {np.mean(r2_scores)}")
    print(f"Average MAE for {target_col}: {np.mean(mae_scores)}")

    plt.xlabel(feature_col)
    plt.ylabel(target_col)
    plt.title(f"{feature_col} vs {target_col}")
    plt.legend()
    plt.show()
```

D Apendix: linear regression 2

```
import numpy as np
import pandas as pd
from scipy.optimize import minimize
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import KFold

def model(params, features):
    a, b, c, d, x = params
    series, strategy, cabins, cgt = features.T
    prediction = a * series + b * strategy + c * cabins + d * cgt + x
    return prediction

def mse(params, features, targets):
    predictions = model(params, features)
    return mean_squared_error(targets, predictions)

data = pd.read_csv("Historic_vessel_data.csv")
features = data[["SERIES", "Building strategy", "Cabins", "CGT"]].values
targets = data["Production outfitting"].values

initial_params = [1.0, 1.0, 1.0, 1.0, 1.0]
bounds = [(None, None)] * len(initial_params)

# 5-Fold Cross validation
kf = KFold(n_splits=5, shuffle=True, random_state=42)
mse_values, r2_values = [], []

for train_index, test_index in kf.split(features):
    features_train, features_test = features[train_index], features[test_index]
    targets_train, targets_test = targets[train_index], targets[test_index]

    result = minimize(mse, initial_params, args=(features_train, targets_train),
    bounds=bounds, method='L-BFGS-B')
    optimized_params = result.x
```

```
    predictions_train = model(optimized_params, features_train)
    predictions_test = model(optimized_params, features_test)

    mse_values.append(mean_squared_error(targets_test, predictions_test))
    r2_values.append(r2_score(targets_test, predictions_test))

print(f"Average MSE: {np.mean(mse_values)}")
print(f"Average R^2: {np.mean(r2_values)}")1
```

E Apendix: Random forest

```
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, r2_score, make_scorer
from sklearn.model_selection import cross_val_score, KFold

# Load data
data = pd.read_csv('Historic_vessel_data.csv')

features = data[['SERIES', 'Building_strategy', 'Powerplant_index',
'CGT', 'OUTFITWEIGTH (w_topside)']]
target_cols = ["Total_hours", "Outfitting_yard", "Project management",
"Production_scaffolding", "Production_outfitting", "Piping & machinery", "Engineering"]

# Create and train the random forest regressor
rf_regressor = RandomForestRegressor(n_estimators=100, random_state=42)

# Perform 5-fold cross-validation
kf = KFold(n_splits=5, random_state=42, shuffle=True)

for target_col in target_cols:
    targets = data[target_col]
    mae_scores = cross_val_score(rf_regressor, features, targets, cv=kf,
    scoring=make_scorer(mean_absolute_error))
    r2_scores = cross_val_score(rf_regressor, features, targets, cv=kf,
    scoring=make_scorer(r2_score))

    print(f"Cross-validated MAE for {target_col}: {np.mean(mae_scores)}")
    print(f"Cross-validated R^2 for {target_col}: {np.mean(r2_scores)}")

    # Fit the model and calculate feature importances
    rf_regressor.fit(features, targets)
    importances = rf_regressor.feature_importances_
    for name, importance in zip(features.columns, importances):
        print(f"Feature '{name}' importance for {target_col}: {importance}")

    print("\n")
```

F Apendix: Cluster-then-regress 2 clusters

```python
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn.model_selection import KFold
import matplotlib.pyplot as plt

# Read data
data = pd.read_csv('Historic_vessel_data.csv')
#data['Total hours'] = data['Total hours'].str.replace(',', '').astype(float)

# Select the Series and Building strategy features
X_clustering = data[['SERIES', 'Building_strategy']]

# Find the optimal number of clusters
n_clusters = 3
kmeans = KMeans(n_clusters=n_clusters, random_state=42).fit(X_clustering)

# Print the cluster centroids
print("Cluster centroids:")
for i, centroid in enumerate(kmeans.cluster_centers_):
    print(f"Cluster {i + 1}: SERIES = {centroid[0]}, Building_strategy = {centroid[1]}")

# Assign the cluster labels to the data
data['Cluster'] = kmeans.labels_

# Create a scatter plot for visualization
fig, ax = plt.subplots()

# Print the rules for each cluster
for i in range(n_clusters):
    cluster_data = data[data['Cluster'] == i]
    print(f"\nCluster {i + 1}:")
    print(f"Building_strategy range: {cluster_data['Building_strategy'].min()}
    - {cluster_data['Building_strategy'].max()}")
    print(f"SERIES range: {cluster_data['SERIES'].min()}
    - {cluster_data['SERIES'].max()}")

    # Linear regression for the most important feature within the cluster
    X = cluster_data[['CGT']]
    y = cluster_data['Total_hours']

    if len(X) >= 2:
        kfold = KFold(n_splits=min(5, len(X)), shuffle=True, random_state=42)
        r2_scores = []
        for train_index, test_index in kfold.split(X):
            X_train, X_test = X.iloc[train_index], X.iloc[test_index]
            y_train, y_test = y.iloc[train_index], y.iloc[test_index]
            lr = LinearRegression().fit(X_train, y_train)
            y_pred = lr.predict(X_test)
            r2_scores.append(r2_score(y_test, y_pred))

        r2 = np.mean(r2_scores)
    else:
        print("Not enough samples for regression analysis.")
```

```
        continue

    # Print the linear regression formula and cross-validated R^2 for the cluster
    print(f"Regression formula: Total hours = {lr.coef_[0]} * CGT + {lr.intercept_}")
    print(f"Cross-validated R^2: {r2}")

    # Plot the datapoints for the cluster
    ax.scatter(X, y, label=f"Cluster {i + 1}")

    # Plot the regression line for the cluster
    x_line = np.linspace(X['CGT'].min(), X['CGT'].max(), 100)
    y_line = lr.coef_[0] * x_line + lr.intercept_
    ax.plot(x_line, y_line, label=f"Regression Line Cluster {i + 1}")

# Set axis labels and legend
ax.set_xlabel('CGT')
ax.set_ylabel('Total hours')
ax.legend()

# Show the plot
plt.show()
```

G Apendix: Cluster-then-regress 1 cluster

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_absolute_error
from sklearn.model_selection import KFold
import matplotlib.pyplot as plt

# Read data
data = pd.read_csv('Historic_vessel_data.csv')

# Select the SERIES feature for clustering
X_clustering = data[['SERIES']]

# Find the optimal number of clusters
n_clusters = 3
kmeans = KMeans(n_clusters=n_clusters, random_state=42).fit(X_clustering)

# Print the cluster centroids
print("Cluster centroids:")
for i, centroid in enumerate(kmeans.cluster_centers_):
    print(f"Cluster {i + 1}: SERIES = {centroid[0]}")

# Assign the cluster labels to the data
data['Cluster'] = kmeans.labels_

# Create a scatter plot of the data points
colors = ['blue', 'green', 'red']
for i in range(n_clusters):
    cluster_data = data[data['Cluster'] == i]
```

```
    plt.scatter(cluster_data['CGT'], cluster_data['Engineering'],
    c=colors[i], label=f'Cluster {i + 1}')

# Initialize lists to store the R^2 and MAE values
r2_values = []
mae_values = []

# Print the rules for each cluster and perform linear regression
for i in range(n_clusters):
    cluster_data = data[data['Cluster'] == i]
    print(f"\nCluster {i + 1}:")
    print(f"SERIES range: {cluster_data['SERIES'].min()}
    - {cluster_data['SERIES'].max()}")

    if len(cluster_data) < 2:
        print("Not enough samples for regression analysis.")
        continue

    # Prepare for cross-validation
    kfold = KFold(n_splits=5, shuffle=True, random_state=42)
    r2_scores = []
    mae_scores = []
    for train_index, test_index in kfold.split(cluster_data[['CGT']],
    cluster_data['Engineering']):
        X_train, X_test = cluster_data[['CGT']].iloc[train_index],
        cluster_data[['CGT']].iloc[test_index]
        y_train, y_test = cluster_data['Engineering'].iloc[train_index],
        cluster_data['Engineering'].iloc[test_index]
        lr = LinearRegression().fit(X_train, y_train)
        y_pred = lr.predict(X_test)
        r2_scores.append(r2_score(y_test, y_pred))
        mae_scores.append(mean_absolute_error(y_test, y_pred))
    r2_values.append(np.mean(r2_scores))
    mae_values.append(np.mean(mae_scores))

    # Print the linear regression formula, R^2, and MAE for the cluster
    print(f"Regression formula: Engineering = {lr.coef_[0]} * CGT + {lr.intercept_}")
    print(f"Avg. R^2 from 5-Fold Cross Validation: {np.mean(r2_scores)}")
    print(f"Avg. MAE from 5-Fold Cross Validation: {np.mean(mae_scores)}")

    # Plot the regression line for the cluster
    x_line = np.linspace(cluster_data['CGT'].min(), cluster_data['CGT'].max(), 100)
    y_line = lr.coef_[0] * x_line + lr.intercept_
    plt.plot(x_line, y_line, c=colors[i], linestyle='--', label=f'Regression Line {i + 1}')

# Calculate and print the average R^2 and MAE value
average_r2 = np.mean(r2_values)
average_mae = np.mean(mae_values)
print(f"\nAverage R^2 value for all clusters: {average_r2}")
print(f"Average MAE value for all clusters: {average_mae}")

# Configure plot labels and legend
plt.xlabel('CGT')
plt.ylabel('Engineering')
plt.legend()
plt.show()
```

H Apendix: Sequentially Restricted Regression

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt

def predict(row, df, a, b):
    series = row["SERIES"]
    cgt = row["CGT"]

    if series == 1:
        return a * cgt + b
    elif series == 2:
        vessel_category = row["VESSELCATEGORY"]
        filtered_df = df[df["VESSELCATEGORY"] == vessel_category]
        avg_ratio = (filtered_df["Total_hours"] / filtered_df["CGT"]).mean()
        return avg_ratio * cgt
    elif series == 3:
        design = row["DESIGN"]
        filtered_df = df[df["DESIGN"] == design]
        avg_ratio = (filtered_df["Total_hours"] / filtered_df["CGT"]).mean()
        return avg_ratio * cgt

def linearonSeriesone(df):
    # Filter the data based on "SERIES" = 1
    a=0
    b=0
    filtered_data = df[df["SERIES"] == 1]

    # Extract the "Outfitting yard" (y) and "CGT" (x) columns
    y = filtered_data["Total_hours"]
    x = filtered_data[["CGT"]]

    # Perform linear regression
    reg = LinearRegression()
    reg.fit(x, y)

    # Display the results
    a = reg.coef_[0]
    b = reg.intercept_
    return a, b

def main():
    a=0
    b=0
    df = pd.read_csv("Historic_vessel_data.csv")
    df = df.dropna(subset=["SERIES", "CGT", "VESSELCATEGORY", "Total_hours", "DESIGN"])
    a, b = linearonSeriesone(df)
    train_df, test_df = train_test_split(df, test_size=0.2, random_state=42)

    # Display the indices of the training and testing sets
    print(f"Training set lines: {train_df.index.tolist()}")
    print(f"Testing set lines: {test_df.index.tolist()}")
```

```python
    predictions = []
    for _, row in test_df.iterrows():
        hours = predict(row, train_df, a, b)
        predictions.append(hours)

    r2 = r2_score(test_df["Outfitting yard"], predictions)
    mae = mean_absolute_error(test_df["Outfitting yard"], predictions)
    print(f"R^2 score: {r2}")
    print(f"MAE: {mae}")


if __name__ == "__main__":
    main()
```

NTNU

Norwegian University of
Science and Technology