Master's thesis

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

Armand Haugerud
Simon Erik Tordhol

# Material Handling in Libraries: Exploring the Impact of Incoming Book Flow and the Potential of Predictive Analytics

## A Case Study on Trondheim Public Library

Master's thesis in Engineering and ICT
Supervisor: Fabio Sgarbossa
June 2023

NTNU
Norwegian University of
Science and Technology

Armand Haugerud
Simon Erik Tordhol

# Material Handling in Libraries: Exploring the Impact of Incoming Book Flow and the Potential of Predictive Analytics

A Case Study on Trondheim Public Library

Master's thesis in Engineering and ICT
Supervisor: Fabio Sgarbossa
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 project is a Master's thesis in ICT and Production Management as a part of the Engineering and ICT study program, at the Department of Mechanical and Industrial Engineering at the Norwegian University of Science and Technology (NTNU). The project is a part of the SmartLIB project, which aims to incorporate technological solutions to improve the operations of Trondheim Public Library. The main topics of the thesis, namely logistics and digitization, correspond greatly with our study program, and the combination of the two fields is of specific interest to both of us.

We would like to thank our supervisor, Fabio Sgarbossa, professor at NTNU, for his valuable feedback and motivation throughout the semester.

We would also like to thank the library staff at TPL, especially Bjørn Tore Nyland, Mildrid Liasjø, and Arild Sørheim, for having been incredibly helpful throughout the semester.

Lastly, we would like to thank our office mates and colleagues, for contributing with valuable reflections and humour throughout the semester.

Trondheim, June 2023

**Armand Haugerud and Simon Tordhol**

# Sammendrag

Denne oppgaven undersøker hvordan materialhåndteringen i et bibliotek påvirkes av den innkommende bokflyten og hvordan prediktiv analyse kan forbedre planleggingen av materialhåndtering. Oppgaven er en del av SmartLIB-prosjektet som ble igangsatt som følge av den økende etterspørselen etter bibliotektjenester som har ført til at Trondheim folkebibliotek har ønsket et «smartere» bibliotek. Målet med prosjektet er å utforske teknologiske løsninger i biblioteksektoren for å løse problemer ved dagens system. Denne artikkelen bygger på resultater funnet i et spesialiseringsprosjekt som identifiserte det å predikere den innkommende bokstrømmen på biblioteker som et aspekt som kan være fordelaktig for bibliotekdriften. Hensikten med artikkelen oppnås ved å svare på følgende forskningsspørsmål:

1. Hvordan påvirker innkommende bokflyt materialhåndtering i et bibliotek?
2. Hvordan kan et bibliotek muliggjøre mer effektiv materialhåndtering?
3. Hvordan kan prediktiv analyse brukes i en biblioteksetting for å forbedre planleggingen av materialhåndtering?
4. Hvordan kan et bibliotek forbedre datakvaliteten for å gjøre prediktiv analyse mer anvendelig?

Den teoretiske delen av denne artikkelen er sammensatt av teori som beskriver materialhåndtering og prediktiv analyse, og er ment å utfylle empiriske resultater oppnådd gjennom casestudiet. Resultatene ble utledet basert på informasjon samlet inn fra møter, semistrukturerte intervjuer, observasjoner og dataanalyse. Til slutt ble prediktive modeller brukt for predikere fremtidige utfall.

Casestudien identifiserte omfattende manuelt arbeid knyttet til materialhåndteringsprosessene for å plassere materiale på midlertidige og originale hyller. Dette ble observert påvirket av hvordan bøkene ble sortert i sorteringsmaskinen. Resultatene indikerte at et bibliotek kunne redusere materialhåndteringen ved å endre sorteringspolicyene for å unngå unødvendige materialhåndteringsoppgaver. Resultatene viste imidlertid at maskinlæringsmodeller ikke var i stand til å predikere den innkommende bokflyten nøyaktig på grunn av at dataen ikke var egnet for formålet.

Oppgaven konkluderer med at innkommende bokflyt i stor grad påvirker manuelle oppgaver knyttet til materialhåndtering på et bibliotek. Videre indikerte resultatene at bruk av prediktiv analyse for å predikere den innkommende bokflyten ikke var fordelaktig for TPL, siden det var unødvendig komplekst sammenlignet med fordelene det ga.

# Abstract

This paper examines how the material handling in a library is affected by the incoming book flow and how predictive analytics can improve the planning of material handling. The study is part of the SmartLIB project which was initiated as a result of the increasing demand for library services which has led to Trondheim Public Library desiring a “smarter” library. The project aims to exploit technological solutions in the library sector to solve problems with the current system. This paper builds upon results found in a specialization project which identified predicting the incoming book flow at libraries as an aspect that could benefit library operations. The objective of the paper is accomplished by answering the following research questions:

1. How does incoming book flow affect material handling in a library?
2. How can a library enable more efficient material handling?
3. How can predictive analytics be applied in a library setting to enhance the planning of material handling?
4. How can a library enhance data quality to make predictive analytics more applicable?

The theoretical part of this paper is composed of theory describing material handling and predictive analytics, and is meant to complement empirical results attained through a case study. Results were derived based on information gathered from meetings, semi-structured interviews, observations, and data analysis. Finally, predictive models were utilized to obtain results of future outcomes.

Case study methods identified extensive manual work related to the material handling processes of placing material on temporary- and original shelves. This was observed to be affected by how the books were sorted in the sorting machine. Results indicated that a library can decrease material handling by altering the sorting policies to avoid excessive material handling tasks. However, results showed that machine learning models were unable to accurately predict the incoming book flow due to a lack of fitness for purpose in the input data.

The paper concludes that incoming book flow affects manual tasks related to material handling at a library to a great extent. Further, results indicated that utilizing predictive analytics to predict the incoming book flow was not advantageous for TPL, as it was unnecessarily complex compared to the benefits it yielded.

***Keywords***— material handling; public libraries; predictive analytics; data analytics; machine learning; data quality

# Table of Contents

# List of Figures

# List of Tables

# List of Acronyms

$R^2$ R-Squared.

**AdaBoost** Adaptive Boosting.
**ANN** Artificial Neural Network.

**DT** Decision Tree.

**GBDT** Gradient Boosted Decision Trees.

**LMS** Library Management System.

**MAE** Mean Average Error.
**MAPE** Mean Average Percentage Error.
**MLP** Multi-Layer Perceptron.
**MSE** Mean Squared Error.

**RF** Random Forest.
**RMSE** Root Mean Squared Error.

**SVM** Support Vector Machine.
**SVR** Support Vector Regression.

# 1 Introduction

This chapter will first clarify the research background before the problem is explained, clarifying the issue under investigation. Following this, the scope of the thesis is outlined, before the research questions and related objectives are presented. Finally, the chapter concludes by presenting an overview of the thesis structure, offering a preview of the subsequent sections and their respective contributions to the overall study.

## 1.1 Research Background

This research paper is written as a master's thesis within ICT and Production Management at NTNU under the SmartLIB project. Initiated in April 2022, the SmartLIB project is a collaboration between NTNU and Trondheim Public Library, with the overarching goal of leveraging technological advancements to create a "smart" library. The project's main purpose is to "develop a smart library system that will solve the current limitations". This entails the creation of a smart library planning and control system, a smart digital assistance tool for staff and patrons, and a smart material handling system. Data plays a fundamental role in achieving a smart library. With the advent of enhanced computational capabilities in recent years, the value of large datasets, commonly referred to as big data, has significantly increased. In 2017, the Economist hailed data as the world's most valuable resource and it has since been likened to "the new oil" (E. D. Peterson 2019). However, similar to crude oil, raw data possesses little intrinsic value without appropriate refinement. The ability to extract valuable insights from data is known as data analytics.

As a result of digital transformation, libraries possess extensive datasets related to loan transactions and book collections. The potential of utilizing library data sources for decision-making was already explored in Nutter (1987), which found that the "ability to collect, organize, and manipulate data far outstrips the ability to interpret and apply them". Furthermore, Arlitsch and Newell (2017) emphasizes the importance of acquiring quantitative and analytical skills to comprehend the value of big data and its potential manipulation, visualization, and analysis for enhancing library operations. Existing literature reviews two primary application areas for data analytics in the library sector: book recommendation and book collection management.

Studies by Sitanggang et al. (2010), S.-T. Yang and Hung (2012), Tsuji et al. (2014), and Iqbal et al. (2020) demonstrate the utilization of data mining and machine learning techniques on loan transaction datasets from academic libraries to identify patterns and frequently occurring events. The insights gained proved valuable for providing book recommendations and informing book procurement. In the context of academic libraries, predictive analytics has also been employed to investigate the relationship between students' book loan history and their academic performance. Lian et al. (2016) proposed a supervised dimension reduction algorithm with multi-task learning for collaborative academic performance prediction and library book recommendation. In Deo (2020), the potential of unsupervised learning was reviewed related to operations within a library.

Particularly, it explored how sentiment analysis could be utilized for the faculties to understand the attitude and approach of students in their learning processes. It was argued that the information gained withe sentiment analysis could be used to suggest changes in the library resources for different courses.

Additionally, Silverstein and Shieber (1996) emphasizes the benefits of utilizing predictive analytics in optimizing storage policies in libraries. The paper suggests employing decision trees to determine which books should be stored off-site in a university library. The performance of decision trees is compared with last-use and random choice policies at Harvard College Library, with decision trees yielding only a fifth of the necessary retrievals compared to second-best. In Uppal and Chandwani (2013), the potential of using data mining to aid decision-making in an academic library is emphasized. It is argued that sequential pattern mining will propose libraries with information regarding frequently loaned books, which should be an important factor in book procurement and layout of the library. Furthermore, in Baba et al. (2016), it is argued that collection management should be conducted based on past circulation data, and not static rules, as is the case in typical university libraries. Further, the paper compares models that are based on synchronous obsolescence, with models based on diachronous obsolescence, to predict the usage of books for improving collection management. Results suggest that the two approaches perform similarly when sufficient circulation data is available, but that the synchronous approach performs better in periods with insufficient circulation data.

While existing literature highlights the potential benefits of applying predictive analytics in the library sector, there are still several areas that warrant further robust research. Primarily, the focus of existing studies predominantly centers on academic libraries, with public libraries receiving limited attention. Although there are similarities between public and academic libraries, the user base in academic libraries is typically less diverse, implying potential differences in the applications of predictive analytics. Furthermore, most studies concentrate on utilizing predictive analytics for book recommendations and collection management, particularly book procurement.

This master's thesis builds upon a specialization project conducted in the fall of 2022, which entailed a case study at Trondheim Public Library. The findings of that project identified incoming book flow as a particularly significant aspect to predict, as it triggers various material handling tasks within the library, making it crucial to consider when planning. Consequently, this thesis aims to expand on the investigation of incoming book flow and its impact on library operations, exploring the advantages that a public library can attain through predictions of incoming book flow. Additionally, the thesis will assess the quality of the data accumulated by Trondheim Public Library, evaluating its suitability as input for predictive models.

## 1.2 Problem Motivation

The metropolitan area of Trondheim has approximately 300 000 inhabitants, and 50 000 of these annually use the services of Trondheim Public Library, hereon referred to as

TPL. TPL consists of nine branches, and a prison branch, spread across the Trondheim region, with the main library located in the city centre. TPL boasts an extensive collection, comprising 430 000 books and various other media items. Annually, TPL facilitates approximately one million loans and organizes 1500 events, contributing to its cultural presence. The library services are financed by taxation and are nonprofit. Public access to the library's resources and services is provided free of charge, with only minor fees imposed on patrons for late returns. The library plays a significant role as a cultural hub in Trondheim, appealing to all age groups.

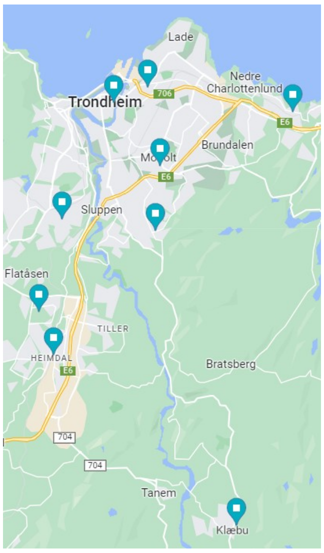


Figure 1: Map of Trondheim with branches mapped.

Based on loan data analysis, it is evident that TPL is experiencing an increase in visitors and loans, nearing the levels observed before the Covid pandemic. Moreover, the number of participants in library events is also on the rise. This growing demand for TPL's services signifies a favorable signal and highlights the importance of further expanding its offerings to cater to the needs of the region's residents. However, at present, TPL faces challenges in meeting the escalating demand due to capacity constraints, and it is unlikely that the funding allocated from the Norwegian National Budget will be increased to a satisfactory level. Consequently, it becomes imperative for TPL to streamline and optimize its processes, particularly in the area of material handling, to minimize manual labor tasks performed by librarians. By doing so, staff members can allocate more of their time and resources to serving patrons and engaging in other value-creating activities such as organizing events.

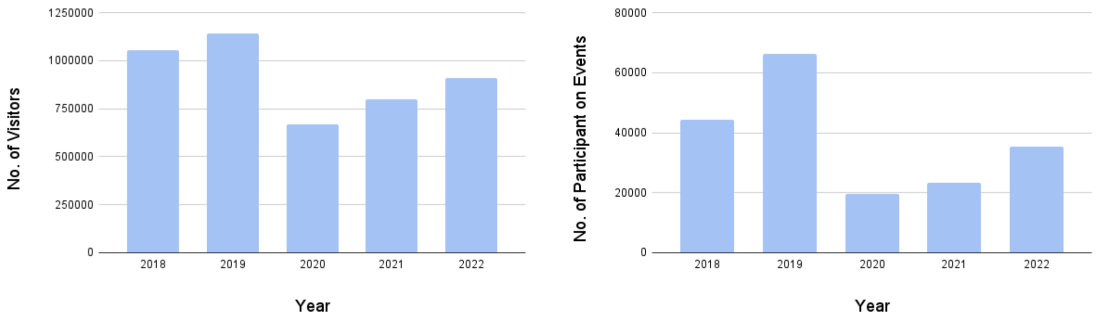


Figure 2: Yearly development of the number of visitors and participants on events.

At present, TPL outsources distribution logistics between its branches. However, the internal movement of books within each library is carried out by the librarians. As the volume of book flow continues to increase, this places a greater burden on staff members, who must allocate more of their time to moving and sorting books within the library. Consequently, the availability of time for serving patrons and engaging in value-adding activities such as organizing events becomes limited. Moreover, these material handling tasks are repetitive, monotonous, and physically demanding, involving heavy lifting, which can have long-term implications for the health and well-being of librarians. The significance of considering human factors and ergonomics in logistic systems is underscored in the research conducted by Sgarbossa et al. (2020). Prioritizing the well-being of library staff, it is desirable to minimize their physical labor.

Given the limited availability of workforce and manhours, it is crucial for the library to obtain an overview of upcoming material flows. This information will enable improved planning and resource allocation, ensuring quality services to patrons every day.

## 1.3 Research Scope

While there are several operational areas that could be explored in relation to predictive analytics, this thesis' scope is confined to investigating how incoming book flow affects library operations, and how predicting the incoming book flow can be beneficial for the library. This will be limited to predicting loan lengths with regression models. The reason why the incoming book flow is deemed particularly interesting is that it initializes several material handling tasks in the library. Further, the thesis will focus solely on material handling related to the inter-library movement of books in TPL's main branch. In this thesis, the term "books" will also include other media that is sorted similarly in the library, such as DVD's and CD's.

## 1.4 Objectives and Research Questions

The overall goal of this thesis is to identify and explore how a library can reduce time spent on manual material handling, with a special focus on applying predictive analytics to obtain information about future incoming book flow. This will be achieved by the following sub-objectives:

1. Map material handling tasks initialized by the incoming book flow.
2. Investigate how much time is spent on manual material handling tasks related to incoming book flow in a day.
3. Identify areas of improvement in the material handling system.
4. Examine how the different sorting setups affect material handling.
5. Investigate the effects of predictive analytics, and determine how applicable predictive analytics is in the current library setting.
6. Develop machine learning models for predicting loan lengths.
7. Explore possibilities that could contribute to improved data quality to better facilitate predictive analytics.

The following table presents the research questions (RQs) that will be examined in this thesis and illustrates how the objectives are mapped to the RQs:

|  | Research Questions | Objectives | Methods | Tools |
|---|---|---|---|---|
| RQ1 | How does incoming book flow affect material handling in a library? | 1. Map material handling tasks initialized by the incoming book flow.<br>2. Investigate how much time is spent on manual material handling tasks related to incoming book flow in a day. | Case Study of case company. Carried out with semi-structured interviews, meetings, and observations. | Microsoft Excel, Stopwatch/timer |
| RQ2 | How can a library enable more efficient material handling? | 3. Identify areas of improvement in the material handling system.<br>4. Examine how the different sorting setups affect material handling. | Case Study of case company. Carried out with semi-structured interviews, meetings, and observations. | Bibliofil, Python, Logs from sorting machine, Microsoft Excel |
| RQ3 | How can predictive analytics be applied in a library setting to enhance the planning of material handling? | 5. Investigate the effects of predictive analytics, and determine how applicable predictive analytics is in the current library setting.<br>6. Develop machine learning models for predicting loan lengths. | Quantitative Data Analysis and Machine Learning Modeling. | Bibliofil, Python, Logs from sorting machine, Microsoft Excel |
| RQ4 | How can a library enhance data quality to make predictive analytics more applicable? | 7. Explore possibilities that could contribute to improved data quality to better facilitate predictive analytics. | Quantitative Data Analysis and Machine Learning Modeling. | Bibliofil, Python, Scikit-learn |

Table 1: Research questions and corresponding objectives

## 1.5 Thesis Structure and Research Outline

The thesis consists of seven chapters. starts with presenting relevant theoretical aspects before a comprehensive case study is conducted, with a focus on material handling tasks related to incoming book flow. Once a thorough understanding of these processes is attained, the case study will explore how the sorting of incoming books can be altered in order to decrease the amount of time spent on material handling tasks. Furthermore, prediction models used for predicting incoming book flow will be presented, alongside the results obtained through these models. Lastly, all findings will be discussed in order to answer the RQs, and possible further work will be proposed.

| | |
|---|---|
| Section 1<br>**Introduction** | Introduction presents the background and motivation of the project, the research scope, the objective and questions, and the structure of the thesis. |
| Section 2<br>**Theoretical Background** | Theoretical background provides a theoretical foundation that is important to establish before starting the case study and predictive modeling. The section looks into material handling, and predictive analytics and gives an introduction to common machine learning models and data quality. |
| Section 3<br>**Methodology** | Methodology presents the approaches used in the research during the project work, both the general methodology for the project and the procedures used for the case study, data exploration and prediction models. |
| Section 4<br>**Case Study** | Case study presents the results that have been used to answer RQ1 and RQ2, constituted by information attained through observations, interviews, and meetings with library staff, in addition to simple data analysis. |
| Section 5<br>**Predictive Modeling** | Predictive modeling presents the procedures and work to answer RQ3 and RQ4 and involves the results from a dynamic sorting experiment, predictive modeling, and analysis of data quality. |
| Section 6<br>**Discussion** | Discussion will summarize and discuss the findings from the case study and predictive modeling with regard to the RQs. |
| Section 7<br>**Conclusion** | Conclusion concludes the paper and discusses the next steps to be done in the project. |

Table 2: Overview of the structure of the thesis

Figure 3 illustrates the research outline for this project. Drawing upon the findings of the previous specialization project and the theory presented in Section 2, four research questions have been formulated to explore the intersection of material handling and predictive analytics in a library context. Additionally, the identification of key outcomes in each research question has played a pivotal role in the shaping of the subsequent research question.

RQ1 and RQ2 stem from the same theoretical foundation, namely material handling in a library setting. RQ1 has predominantly been investigated through a combination of meetings, interviews, and on-site observations. These methods led to the key outcomes highlighted in the figure. RQ2 has primarily been explored via data analysis and simulations aimed at assessing the performance of proposed sorting configurations.

RQ3 establishes a connection between material handling in a library and predictive analytics. The research approach employed to derive key outcomes for addressing this question has also involved data analysis and simulations to ascertain the impact of various sorting setups, in addition to predictive modeling, to review the applicability of predictive analytics in this case. Finally, RQ4 is more related to the theory of predictive analytics. To address this RQ, machine learning techniques have been employed to yield results.

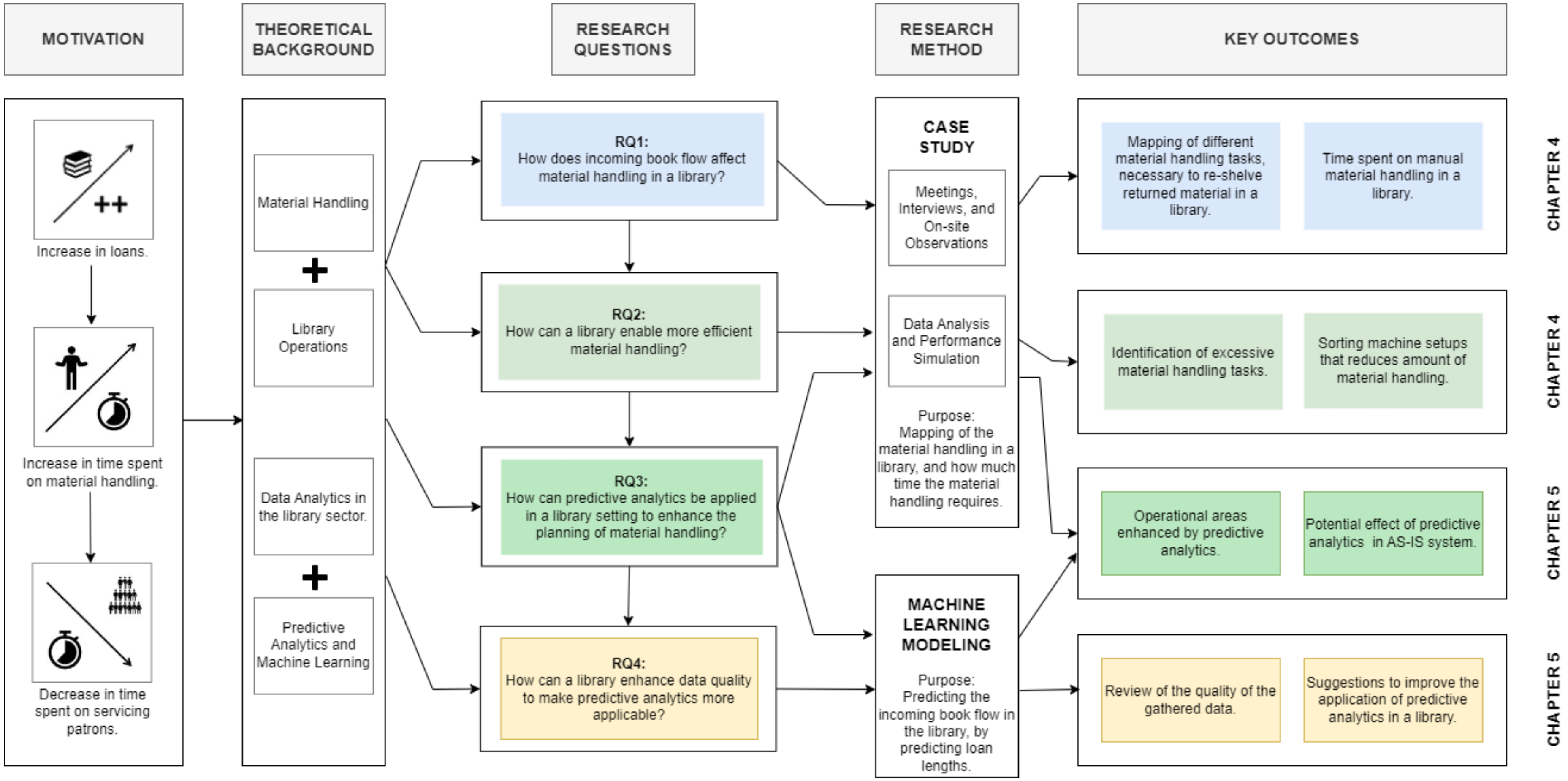

MOTIVATION
THEORETICAL BACKGROUND
RESEARCH QUESTIONS
RESEARCH METHOD
KEY OUTCOMES
Increase in loans.
Increase in time spent on material handling.
Decrease in time spent on servicing patrons.
Material Handling
+
Library Operations
Data Analytics in the library sector.
+
Predictive Analytics and Machine Learning
RQ1: How does incoming book flow affect material handling in a library?
RQ2: How can a library enable more efficient material handling?
RQ3: How can predictive analytics be applied in a library setting to enhance the planning of material handling?
RQ4: How can a library enhance data quality to make predictive analytics more applicable?
CASE STUDY
Meetings, Interviews, and On-site Observations
Data Analysis and Performance Simulation
Purpose: Mapping of the material handling in a library, and how much time the material handling requires.
MACHINE LEARNING MODELING
Purpose: Predicting the incoming book flow in the library, by predicting loan lengths.
Mapping of different material handling tasks, necessary to re-shelve returned material in a library.
Time spent on manual material handling in a library.
CHAPTER 4
Identification of excessive material handling tasks.
Sorting machine setups that reduces amount of material handling.
CHAPTER 4
Operational areas enhanced by predictive analytics.
Potential effect of predictive analytics in AS-IS system.
CHAPTER 5
Review of the quality of the gathered data.
Suggestions to improve the application of predictive analytics in a library.
CHAPTER 5

Figure 3: Visual representation of the research outline.

# 2 Theoretical Background

This chapter aims to introduce key terminologies that hold significance for this thesis. The chapter is structured into two sections. Initially, it begins by explaining material handling, accompanied by related terms such as human factors and ergonomics. Notably, emphasis is placed on material handling within a library context. Subsequently, the chapter proceeds to explicate predictive analytics and associated terminologies. The topics covered in the theoretical background have been selected to explain some of the topics considered important in this thesis and to support the results attained in Section 4 and Section 5.

## 2.1 Material Handling

According to Coyle et al. (1992), material handling refers to the short-distance movement of goods within the confines of a building or between a building and a transportation vehicle. Additionally, Apple 1972 defines material handling as the movement, protection, storage, and control of materials within the supply chain, contributing to the creation of time and place utility. Material handling holds significance in both manufacturing and service industries. Theoretically, material handling is perceived as an additional cost to a product since it does not directly provide form utility, as observed in manufacturing. However, in practice, material handling can add substantial value to a product through the execution of various processes (Kay 2012). Material handling involves the incorporation of manual, semi-automated, and automated equipment and systems to facilitate the efficient functioning of the supply chain (Institute 2023).

### Technological Trends in Material Handling

The introduction of Industry 4.0 has led to a rapid increase in the use of technologies in material handling systems. As a result, of these new technologies, material handling can to a large degree be automated, depending on the industry. First of all, equipment and products are interconnected through Internet of Things (IoT), such as RFID-sensors. The introduction of IoT has greatly increased the transparency and traceability of production processes, and inbound- and outbound logistics, which is crucial to managing the increasing variety of customer orders (C. K. Lee et al. 2018). Further, the emergence of IoT has facilitated the integration of robotics into material handling processes, leading to a reduction in human operators' involvement in monotonous and repetitive tasks, as well as a decrease in material handling time. Robotics, such as automatic guided vehicles (AGVs) and autonomous vehicles, are frequently employed in material handling systems to achieve these objectives. AGVs operate by following pre-defined routes, while autonomous vehicles possess a certain level of intelligence, enabling them to make decisions based on real-time perceptions of their environment.

### Material Handling in a Library

Libraries, as compared to the manufacturing industry, operate with a more service-oriented supply chain (Hye et al. 2019). The provision of high-quality products to customers relies heavily on efficient reverse logistics operations. The most prevalent material handling tasks in a library are constituted of moving or sorting books. These tasks involve shelving, sorting, preparing reservations or interlibrary loans, and moving material to and from the storage areas. Typically, librarians are responsible for performing these material handling duties, although varying degrees of robotic automation have been implemented in numerous libraries. One of the most common automated devices employed is the sorting machine, which automates either a portion or the entirety of the sorting and storage process. The process of moving books around the library is often done manually with trolleys or boxes, but some libraries also have begun incorporating AGVs to facilitate the inter-library movement of books. Oodi, a Finish public library located in Helsinki, is considered one of the frontrunners for technological solutions in the library sector. By integrating automated solutions, librarians can reduce the time spent on material handling, and allocate more time servicing patrons. However, technologies automating material handling are often associated with significant initial costs. Pop and Mailat (2011) emphasizes that a modern library must keep pace with the continuously evolving technological environment. This can often require reorganization and upgrades to the library facilities.

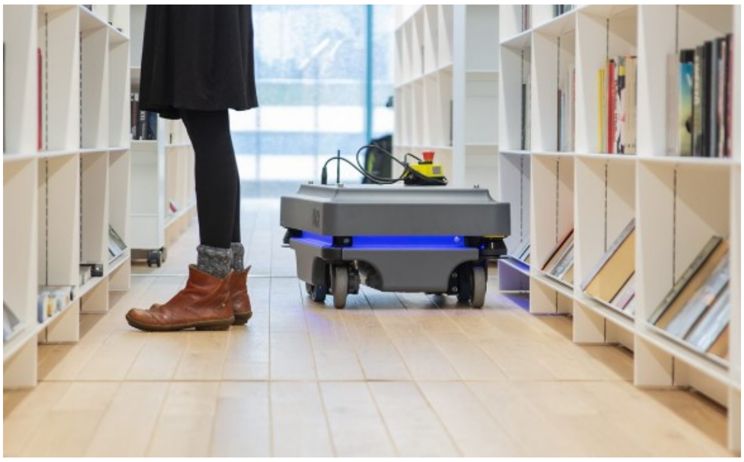


Figure 4: AGV employed at Oodi library, used to transport books between different floors.

Source: Oodi 2019

To ensure that material handling is performed efficiently, RFID technology is widely popular within libraries. The RFID-system is normally composed of tags, readers, and middleware software. In the case of a library, the tags are placed on the books or other material, and the readers are placed at staff workstations and self-return machines. The middleware operates an important function in the system as it supports the communication between the Library Management System (LMS) and the reader (Ayre 2012). An Library Management System is an ERP system specialized for library operations. It is an integral part of the modern library, and stores information related to the book collection, loan transactions, and patrons. This is also the case for TPL, which is currently using Bibliofil, provided by Bibliotek-Systemer AS. An LMS is typically composed of several modules and subsystems, for instance, a cataloging system, acquisition system, loans systems, and management information systems (Ferguson and Hebels 2003). These systems store large amounts of data. An example of a data source is RFID data, which at TPL is generated, for instance, every time a book is passed through the sorting machine. Through Bibliofil, it is possible to extract data describing the book collection at TPL, the loan and return transactions, and check the status of every copy in the collection. It is possible to send out reminders to patrons through Bibliofil, and patrons can also extend loans in the system.

### Human Factors and Ergonomics

An important aspect to consider when reviewing logistics- and manufacturing systems, is how human factors and ergonomics (HF&E) are considered. HF&E is defined as "the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance" (IEA 2000). An ergonomist evaluates the demands of a specific task with reference to the capacity of the workers to perform the task over a certain time period. This is to be able to design the workplace as optimal as possible for a greater part of the workforce (Fernandez 1995). In libraries, material handling can be straining on the employees, as it involves moving and lifting multiple books, often at different heights because of the different levels on the shelves. Vijayakumar et al. (2022) emphasizes the importance of understanding the impact of design and management decisions on the interaction between human and system elements, and it is highlighted as an important performance factor. A result of neglecting or underestimating HF&E is reduced performance and quality of work, and increased cost and time consumption in operations (Sobhani et al. 2017). A successful implementation can lead to improved health and safety of the workers, job satisfaction, improved morale of workers, and a decrease in absenteeism (Fernandez 1995).

The daily operations within a library involve several material handling tasks, exposing employees to potential musculoskeletal strains in their backs, necks, and arms. As previously mentioned, examples of material handling tasks prevalent in a library include sorting, shelving, and moving books. In Thibodeau and Melamut (1995), five factors that contribute to cumulative trauma disorders are presented, which involve elements like repetition, forceful exertions, and unfavorable posture - attributes that are notably inherent in material handling tasks in a library setting. Consequently, considering ergonomic principles and human factors in the design of a library's material handling system is crucial, as it not only decreases the risk of such injuries but also enhances employee well-being, thereby resulting in long-term efficiency gains. Moreover, as explained in Labajo (2017), librarians are essential for providing quality information services to patrons. The quality of these services will likely be affected by the librarians' well-being, magnifying the importance of the inclusion of ergonomics and human factors in the design of material handling systems in libraries. In addition to tailoring the material handling system to accommodate employee needs, it is imperative for librarians to be cognizant of the ergonomic challenges inherent in their daily work routines, thereby avoiding excessive strain and awkward lifting.

## 2.2 Predictive Analytics

Predictive analytics concerns a variety of statistical algorithms and machine learning techniques, which aim to uncover relationships and patterns within large volumes of data, that can be used to predict behavior and events (Eckerson 2007). Predictions of future outcomes are generally considered as having high business value, but developing prediction models is also considered labor-intensive and time-consuming. advanced computer processing and database technologies have significantly enhanced the effectiveness of predictive analytics. Traditional statistical techniques, such as linear- and logistic regression are now complemented with techniques such as neural networks and genetic algorithms. These algorithms excel in handling larger datasets with numerous variables and exhibit greater resilience to anomalies and noise. Consequently, predictive analytics has emerged as a vital tool across diverse sectors, facilitating improved decision-making and strategic planning.

Predictive analytics is commonly applied in order to obtain estimates of the length of different processes, such as order deliveries. To maintain competitiveness companies need to promise relatively short delivery dates, while also ensuring that the promised date is met. Specifying the delivery date in the contract gives both the seller and the customer a date to deal with (Hiroyuki 2012). Examples include delivery date estimation for postal services, production lead times, and travel-time predictions. The complexity of a time prediction problem can, however, greatly vary depending on the problem. The more factors affecting the process, the more uncertainty will be related to the date estimation (Seyedan and Mafakheri 2020). For example, it is easier for an Amazon customer to receive a precise delivery time estimate on their order compared to a buyer ordering a new-built engineer-to-order ship. The value chain of an ETO shipbuilder becomes complex as a result of limited data on custom components and can involve many different actors which makes it hard to make a precise time estimation. In almost all industries, predicting the time related to a process, either internal or external, is of great relevance, as it allows for better planning and customer satisfaction. In a library setting, the prediction of, for instance, loan lengths would contribute to obtaining information about future incoming book flow. This could be utilized to forecast the demand for library services, thereby providing decision support in relation to capacity planning.

### Statistical Modeling and Algorithms

Statistical modeling is a sub-branch of predictive analytics, and can be considered the frontrunner of machine learning. A statistical model is by McCullagh (2002) defined as a set of probability distributions on the sample space **S**. Another description of statistical modeling is the mathematical relationship between random and non-random variables (Adèr 2008). Statistical modeling is commonly used for predictive and forecasting tasks based on existing data and is often applied to the same problems as machine learning models (Ij 2018). This is logical since machine learning is built upon the theory of statistics. The overall goal of both sets of models is to improve the accuracy while simultaneously minimizing a loss function, thus making few and small errors. The difference lies in how the loss is minimized, the machine learning models use non-linear algorithms while the statistical models use linear processes (Makridakis et al. 2018).

## 2.3 Machine Learning

Machine learning represents a prominent domain within computer science and predictive analytics, often discussed alongside artificial intelligence, deep learning, and data mining. While these terms are occasionally used interchangeably in practice, they possess distinct meanings (Jordan and Mitchell 2015). This paper's main focus will be on machine learning and deep learning, a subset of artificial intelligence. Mitchell and Mitchell (1997) defines machine learning as "the study of computer algorithms that allow computer programs to automatically improve through experience". In essence, machine learning involves methods and systems that are able to learn and adapt autonomously, without the need for additional human intervention. The basis of machine learning models is mathematical models.

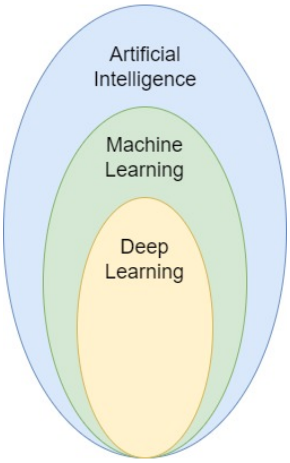


Figure 5: Relationship between artificial intelligence, machine learning and deep learning

Source: Sindhu et al. 2020

It is normal to divide the learning process of a machine learning model into two groups; supervised and unsupervised learning. This thesis will focus on supervised learning.

### 2.3.1 Supervised Learning

Supervised learning refers to a learning process where labeled data is used during the training of a machine learning model. The model bases its predictions on test data according to the labels observed in the training data. The model adjusts its accuracy with the use of loss functions until the error is minimized, or reaches a predetermined threshold. The performance and results of the supervised learning models can be evaluated by the use of different scoring and error metrics (IBM 2020). Supervised learning can be divided into two problems, regression, and classification. This thesis will only focus on regression, as this is what is most suited to the examined prediction problem.

A regression model aims at predicting a continuous value as close as possible to the actual value. The model uses a mathematical approach for predicting continuous outcomes and adjusts its predictions based on learning from the training data. The final regression model generally involves plotting a line that fits the data points in the best way possible. Regression models are commonly used to perform forecasting and time series prediction (Huang et al. 2020). For evaluating regression models, the most common scoring metrics are Mean Squared Error, Root Mean Squared Error, Mean Average Error, Mean Average Percentage Error and R-Squared (Botchkarev 2019).

### 2.3.2 Steps of Machine Learning

The process of applying machine learning models consists of several steps to allow the algorithms to best analyze and learn from existing data, before predicting future outcomes. These steps are illustrated in the figure below:

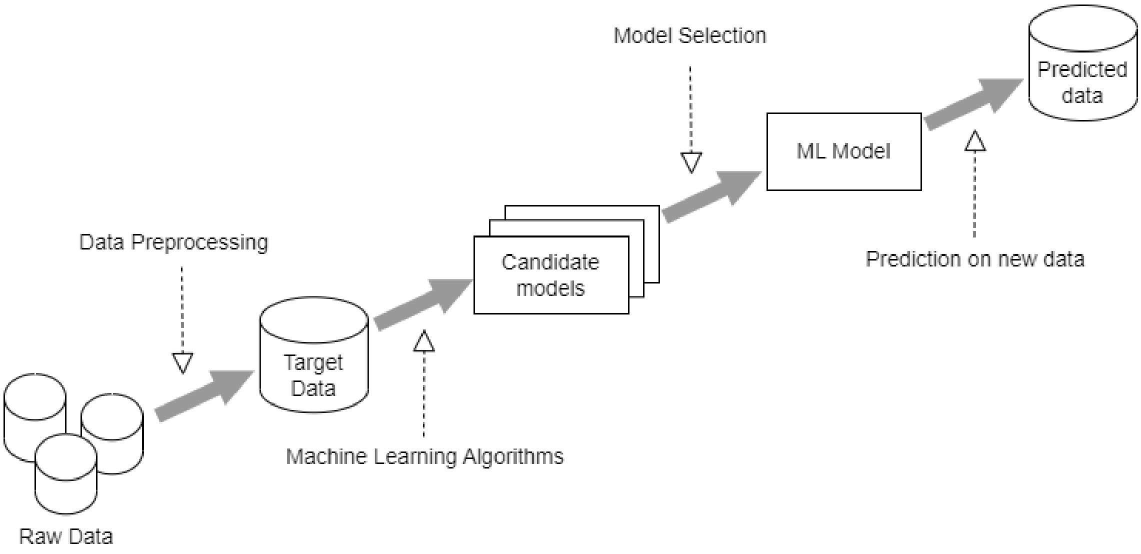


Figure 6: Machine Learning Process from raw data to predictions

The subsequent paragraphs will provide further explanations of the different process steps.

### 2.3.3 Data Preprocessing

The foundation for all machine learning techniques is large volumes of data. The quality of the input data directly affects the machine learning model's ability to learn and later predict. Preparation and exploration of the input data is an integral part of building a machine learning model. Data preprocessing makes it possible to adapt input data to the requirements posed by different models, enabling to process data that would otherwise be infeasible (Garcia et al. 2016). Data preprocessing is normally divided into four main parts (Mushtaq 2019):

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

### Data Cleaning

Data cleaning is considered the initial step for ensuring high-quality input data for the machine learning models and is regarded as the first step of preprocessing. Poor data quality in the context of machine learning can lead to poor decision-making and unreliable analysis (Gudivada et al. 2017). The problem of detecting and repairing *dirty* data is considered one of the perennial challenges in data analytics (Chu et al. 2016), as the presence of errors, missing values, inconsistency, incompleteness, outliers, and noise in the data can lead to poor data quality (G. Y. Lee et al. 2021).

An outlier is a value in a dataset that vastly differs from the other values, which will usually affect the mean in the dataset. These values could be valid data points or noise. Noisy data is data that holds meaningless information which makes it hard for the user or system to understand and interpret the data correctly (Klimushkin et al. 2010). Therefore, the presence of noisy data and outliers can affect the performance of a machine learning model in an undesirable way. When noise is present in the data set, the machine learning might interpret this data as patterns, hence resulting in a less accurate description of the target concept and thus be of a lower utility. Overfitting is a common problem appearing when a model learns from outliers and noisy data. There are different techniques used for avoiding overfitting and the purpose is to make sure that the parameters created by the model do not get too complex in order to fit this noise and outliers (Teng 1999, Clark and Niblett 1989). It is a common procedure to remove the noisy data and outliers from the dataset based on certain evaluation mechanisms (John 1995, Brodley, Friedl et al. 1996).

The occurrence of missing values is a common problem when working with data analysis. A missing value occurs when an empty data value is stored in a variable of an observation (Fernstad 2019). Human error, machine error due to malfunction, and the merging of unrelated data are examples of what could cause a missing value (Emmanuel et al. 2021). There are different techniques for handling these missing values, where deletion of instances or replacement with estimated values are among the most common techniques (Ludbrook 2008).

### Data Integration

In the process of data collection, it is common to gather information from various sources, leading to the possibility of redundancy and inconsistency within the merged dataset. To address this issue and ensure data integrity, data integration techniques are employed. One such technique involves manual integration, although this approach is time-consuming and viable only for smaller datasets. Alternatively, data warehousing provides a solution for effective data integration.

### Data Transformation

Following data integration, the next step is data transformation, which aims to convert the data into a suitable format for modeling purposes.The rationale for data transformations can be categorized into two types: mandatory transformations and optional quality transformations. Mandatory transformations are necessary to ensure data compatibility, such as converting non-numeric features into numeric representations and resizing inputs to a fixed size (Learning 2022b). QOn the other hand, optional quality transformations are not always essential but can enhance model performance. Through data transformation, values within the dataset are adjusted or modified to enable the model to establish optimal learning parameters, resulting in improved accuracy (Kusiak 2001). Various methods are available for implementing data transformation on the dataset, including filling in missing values, feature discretization, feature normalization, feature generalization, feature specialization, and feature engineering (Al Shalabi et al. 2006, Obaid et al. 2019).

Within a dataset, values can be categorized as either categorical or numerical. Numerical values, in turn, can be further classified as either continuous or discrete. In situations where numerical values exhibit a discrete nature, such as postal codes, it becomes necessary to treat them as categorical variables. This approach prevents the model from erroneously attempting to establish a numeric relationship between postal codes, instead enabling the creation of distinct signals for each individual postal code (Learning 2022b).

**Feature Normalization:** The goal of feature normalization is to scale the attribute so that the values fall into a specified range (Obaid et al. 2019). Normalizing the features can both improve accuracy and runtime for machine learning models and additionally avoid errors. There are different techniques and models used for normalization, for instance, scaling to a range, clipping, log scaling, and z-score (Learning 2022c).

**Feature Discretization:** Discretization is the process where continuous variables are converted into discrete ones by splitting into a finite number of sub-ranges called intervals, buckets, or bins (Cebeci and Yildiz 2017). There are different techniques used for discretization, where the binning techniques are the most common (Dougherty et al. 1995). Binning is done by using equally spaced boundaries or quantile boundaries. Equally spaced boundaries can lead to uneven distribution of points among the buckets while with quantile boundaries each bucket has the same number of points. The two methods can generate completely different results based on how the machine learning model models the feature (Learning 2022a).

**Feature Engineering:** Feature engineering is regarded as an important but labor-intensive component of the development of a machine learning model. Zheng and Casari 2018 define feature engineering as “the process of formulating the most appropriate features given the data, the model, and the task”. By defining new features in a data set, the goal is to decrease computational complexity in addition to increasing the learning effect (Cai et al. 2018). accuracy. These new features might be ratios, differences, or other mathematical transformations or combinations of existing features (Heaton 2016).

### Data Reduction

Training a machine learning model on a large data set can be computationally complex. It is therefore important to find a balance between accuracy and computational complexity (Cunningham 2008). Furthermore, different features in the data sets are characterized with varying degrees of importance regarding the accuracy of the model, and some might even degrade the overall performance of the model. This phenomenon is popularly referred to as the “curse of dimensionality” (Verleysen and François 2005). To avoid this, data reduction is performed, which can be described as the process of reducing the number of variables in the data without losing validity. The most common approaches used for data reduction are feature selection, feature extraction, and feature transformation. Feature extraction is based on combining existing features in the dataset, before removing the original features.

**Feature Transformation:** Feature transformation is a technique that combines multiple features into a single new feature. The most common method for feature transformation is Principal Component Analysis (PCA) (Cunningham 2008). PCA aims to extract the most important information from the dataset while simultaneously reducing its dimensions without sacrificing vital data. This objective is accomplished by generating new variables known as principal components, which are essentially linear combinations of the original variables (Abdi and Williams 2010).

**Feature Selection:** Feature selection aims to identify the “best” subset of original features, rather than transforming the data into entirely new dimensions (Cunningham 2008). There are different approaches and criteria used for feature selection, and examples include removing features with low variance, univariate statistical tests, such as an F-test, and recursive feature elimination (Pedregosa et al. 2011). Furthermore, bivariate statistical tests are often applied in the form of correlation analysis for feature selection. Through correlation analysis, it is possible to determine the strength of the relationship between two item sets, usually expressed as a decimal number (Kumar and Chong 2018). Machine learning models will perform better if features with low correlation to the target variable are excluded from the dataset, both computationally and in terms of accuracy (Zheng and Casari 2018). However, amongst the independent variables in a dataset, or features, it is beneficial with low correlation. Non-correlated features in the data improve the learning rate of the model, interpretability will be higher and the data will have less bias.

Correlation is usually computed through Pearson’s correlation coefficient. Considering the variables X and Y, the correlation can be calculated as:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{1}$$

The coefficient ranges from -1 to 1, with -1 meaning perfect negative correlation, and 1 meaning perfect positive correlation. Zero equals no correlation. Typically, you set a minimum threshold, and remove the features with correlation below the given threshold.

### 2.3.4 Learning Algorithms

After the data has been preprocessed, it is time to introduce machine learning algorithms to the data set. Deciding which algorithm to use is a difficult task, as the performance of machine learning algorithms is heavily case-dependent. “No free lunch” is a theorem describing the impossibility of a universal optimization strategy (Wolpert and Macready 1997), and it is considered applicable also in machine learning.

#### Decision Trees

Decision Tree (DT) is a tree-based supervised learning algorithm commonly used in machine learning. DTs are popular in several application fields because they are easy to interpret (Charbuty and Abdulazeez 2021). However, DTs are known to be prone to overfitting. DTs can be applied in both regression and classification problems. The method follows a common tree structure.

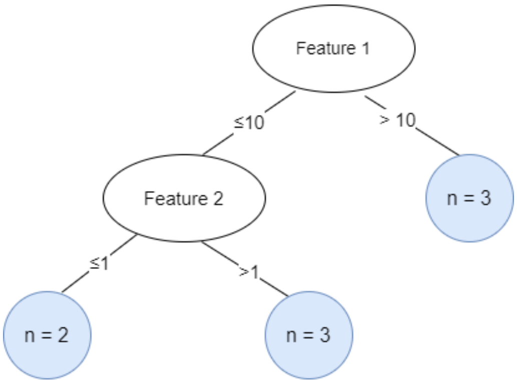


Figure 7: Example of a simple regression tree

#### Support Vector Machine

Support Vector Machine (SVM) is another supervised machine learning model that is used for both classification and regression problems. When the algorithm is used for regression, it is referred to as Support Vector Regression (SVR). The objective of an SVM is to find the hyperplane that maximizes the separation of classes of data points (Noble 2006). Points with the least margin to another class are called support vectors. The SVR aims to find the hyperplane which interferes with the most points. Support Vector models are considered particularly effective when working with data in high dimensional spaces (multiple features).

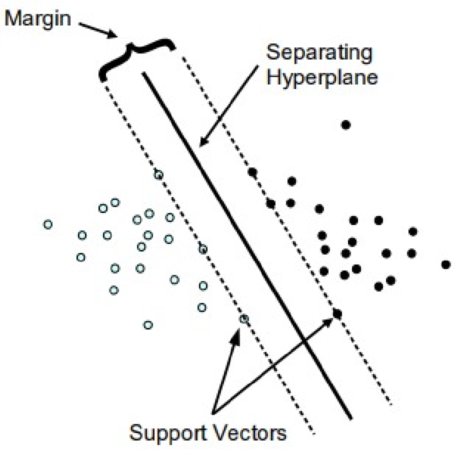


Figure 8: SVM used for binary classification.

Source: Meyer and Wien 2015

### Artificial Neural Network

Artificial Neural Network (ANN) is a deep learning model inspired by the neural networks in the human brain. It is built up by layers of nodes, or neurons, and each layer communicates with the neighborly, fully connected layers. The connections between neurons are weighted by a weight function. A network contains one input- and output layer, and the information is passed from the input layer and processed in the middle, or hidden, layers before the result is presented by the output layer. Many other deep learning models are based on an ANN.

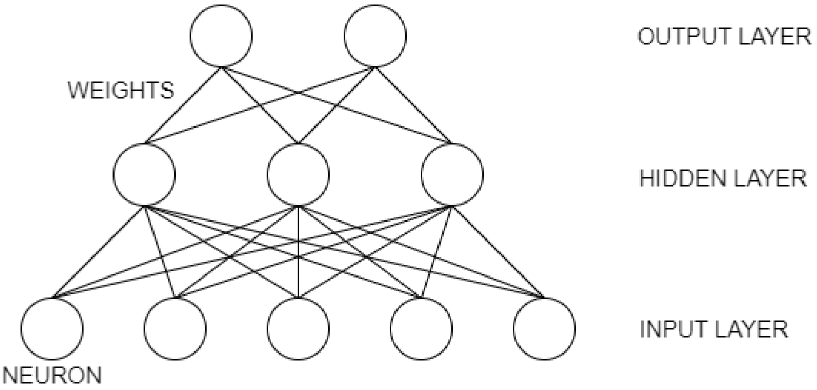


Figure 9: Architecture of an ANN.

Source: S.-C. Wang 2003

### Random Forest

Random Forest (RF) is an ensemble machine learning model combining the outputs from multiple decision trees to one single output(Livingston 2005). The method can be used for both regression and classification problems. In most cases RF is more accurate than DT. What separates RF from DT is the computational complexity of RFs, which makes DTs faster to train. But the complexity also brings the advantage of RFs in most cases being less prone to overfitting since the result from several models is considered and combined(Ali et al. 2012).

### Adaptive Boosting

Adaptive Boosting (AdaBoost) is an ensemble machine learning model. The model combines the output from several algorithms, often referred to as "weak learners", into one single output(Rojas et al. 2009). The output is calculated by weighting the outputs from the weak learners into one strong learner. The weights are adjusted for each iteration when a new weak learner is introduced. Decision Trees are often used to generate weak learners(Feng et al. 2020). AdaBoost can be used for both regression and classification problems.

### Gradient Boosted Decision Trees

Gradient Boosted Decision Trees (GBDT) is an ensemble machine learning model. The model works similarly to AdaBoost where the outputs from several models, Decision Trees, are combined into one single output. Instead of weighting the outputs like in AdaBoost, the gradient boosting method consecutively fits new models based on a loss function in order to build a more accurate model(Natekin and Knoll 2013). The model is used for both regression and classification problems.

### 2.3.5 Model Selection

Further, model selection is performed. Model selection is the process concerned with choosing the best learning method, and its optimal hyperparameters. However, attaining an understanding of what is "best" can be a difficult task, as performance evaluation of machine learning models is not trivial (Raschka 2018). Often there is no "best" model, but some can instead be said to be "good enough". The term "good enough" could imply that a model meets predefined requirements, has a good ratio of complexity and accuracy, or stands out compared to naive models or other machine learning models.

An important pitfall to consider in model selection is overfitting. Overfitting means that the model models too similar to the training data. Mirroring the training data to an excessive degree, even mirroring outliers and irregularities, results in the model not being generalizable to new data, which results in poor performance when working with unseen data (Dietterich 1995). There are different methods that can be used to prevent overfitting, the easiest being training with more input data, as the lack of training data is often the reason for overfitting. If large data sets are not available, other methods dealing with overfitting will need to be applied. Such methods include data augmentation, feature selection, and cross-validation. When using data augmentation, small changes to the sample data are done for each time the model processes it. Cross-validation is a data resampling method used to assess the generalization ability of predictive models (Hastie et al. 2009). In cross-validation, the learning set is randomly partitioned into test and training sets. If using standard k-fold cross-validation, the data will be separated into k parts and run k times with each part as the validation set once (Anguita et al. 2012). Another aspect to consider is imbalanced domain learning where being able to predict the extreme values, or rare cases, is of high priority. Normally, regression models are rated based on overall error estimates, with a focus on minimizing the total error. Usually, this is indicative of model performance given the predominance of cases with target values near the central tendency of the distributions (Ribeiro and Moniz 2020). However, when it is desirable to be able to predict the unusual cases, the best model might not be the one with the lowest overall error, but instead the one with the lowest error to the most important target values.

Another important aspect of model selection is hyperparameter optimization, sometimes referred to as tuning. L. Yang and Shami (2020) defines hyperparameters as “the parameters used to configure a machine learning model, or to specify the algorithm used to minimize the loss function”, and differs based on machine learning algorithms. Hyperparameter tuning can be done either manually, which requires a certain degree of expertise to be effective, or by automatic techniques. As mentioned in the section on data preprocessing, it is important to balance computational complexity and precision. The same applies when tuning hyperparameters, as it can be computationally costly to evaluate the performance of a large set of hyperparameters.

### 2.3.6 Model Performance Metrics

Before selecting the final model, the performance of the models selected in the model selection step will need to be evaluated and compared. It can be hard to evaluate the results of machine learning models just by looking at the predictions. A performance metric quantifies the performance of a model through a function, into a single score. The performance metrics have their advantages and shortcomings. Therefore it is common to combine several metrics to obtain a holistic view of the performance of a model (Hicks et al. 2022).

#### Mean Absolute Error

Mean Average Error (MAE) is a metric used to quantify the errors between paired observations that represent the same phenomenon. It provides a measure of the average absolute difference between the predicted and actual values (Willmott and Matsuura 2005).

#### Mean Absolute Percentage Error

Mean Average Percentage Error (MAPE) is a metric used to measure the average percentage difference between the predicted and actual values. It quantifies the average absolute percentage deviation between the predicted values and the true values, providing an indication of the overall accuracy of the predictions (De Myttenaere et al. 2016).

#### Root Mean Squared Error

Root Mean Squared Error (RMSE) is a metric that calculates the square root of the average squared difference between the predicted values and the actual values. It is a commonly used measure of the overall error or discrepancy between the predicted and observed values. RMSE provides a measure of the typical magnitude of the errors and is often preferred when larger errors have a greater impact (Willmott and Matsuura 2005).

#### R-Squared

R-Squared ($R^2$) is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model (Cameron and Windmeijer 1997). It indicates the goodness of fit of the model to the data and ranges from 0 to 1 (Cameron and Windmeijer 1996). An $R^2$ value of 1 signifies that the model perfectly predicts the dependent variable based on the independent variables, while a value of 0 indicates that the model does not explain any of the variability in the dependent variable.

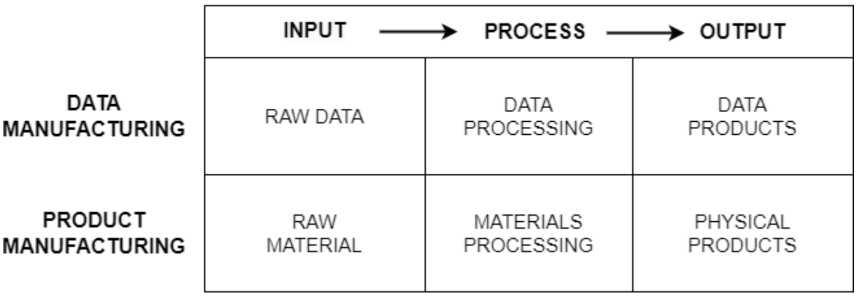
### 2.3.7 Prediction on New Data

After model selection, the final model has been decided, and the model is ready to be implemented, perform predictions, and provide business value. This is the step most end-users see when they use machine learning models in their respective industries. The challenge for end-users is utilizing the predictions to improve operations as much as possible. The model should be continually improved based on feedback.

## 2.4 Data Quality

Data quality is an important term with regard to data analytics. Hazen et al. (2014) compares poor data quality to being thirsty while at sea; you are surrounded by water, but nothing worth drinking. The same degree of uselessness characterizes the consumption of poor-quality data. Further, it is estimated that the costs of poor data quality can be as high as 8% to 12% of revenue for a typical organization.

The importance of data quality from a business perspective is usually understood, but not always dealt with. It is crucial that companies have clear visions and standards for their data "manufacturing". In fact, R. Y. Wang et al. (1995) drew an analogy between the manufacturing of data and the manufacturing of products, explained by a simple framework:



| | INPUT → | PROCESS → | OUTPUT |
|---|---|---|---|
| DATA MANUFACTURING | RAW DATA | DATA PROCESSING | DATA PRODUCTS |
| PRODUCT MANUFACTURING | RAW MATERIAL | MATERIALS PROCESSING | PHYSICAL PRODUCTS |

Source: R. Y. Wang et al. 1995

Without the proper raw materials, a company would never be able to attain products that hold a satisfactory level. Additionally, as stated in Hazen et al. (2014), raw materials are generally depleted during a production process, meaning that the harm of a bad batch will be limited. However, poor data will remain until it is actively removed, and can therefore cause larger damage. This stresses the importance of monitoring the data quality in a company. Furthermore, continuous improvement is desired in production processes, and is what allows companies to continue competing. Jones-Farmer et al. (2014) stresses the importance of a framework to ensure continuous improvement in the data manufacturing processes as well, analogous to the Define, Measure, Analyze, Improve, Control (DMAIC) cycle as ascribed by Six Sigma and frequently used to monitor and improve production processes.

### Determining Data Quality

Moreover, it is commonly accepted that data quality involves multiple dimensions (Pipino et al. 2002), and according to IBM (2023), data quality measures how well a dataset meets criteria for:

- Completeness - The amount of data that is usable or complete.
- Uniqueness - The amount of duplicate data in the dataset.
- Validity - How much of the data matches the required format.
- Timeliness - The readiness of the data within an expected time frame.
- Accuracy - Correctness of the values in the dataset, based on a predetermined “truth”.
- Consistency - How data from different sources measure the same metric, and the relationship between different data sources.
- Fitness for purpose - The extent actually provides advantages for its intended use.

These dimensions describing data quality might be represented by different metrics. For instance, it can be straightforward to measure completeness, accuracy, and uniqueness, as these dimensions can be described with a simple ratio. A typical way of calculating this ratio is by subtracting the ratio of erroneous data to 1. For these ratios to be consistent, Pipino et al. (2002) stresses the importance of having predefined, case-specific criteria for what should be considered accurate or complete. However, dimensions such as consistency and fitness for purpose are not described sufficiently by a ratio, as they are more complex, often case-dependent, and dependent on multiple variables. Fitness for purpose is, for instance, heavily reliant on how the collected data is relevant to the problem, which is defined by the data scientist based on expert estimations (Heinrich et al. 2018). Moreover, fitness for purpose is often not referred to as a dimension, but rather what defines data quality (Tayi and Ballou 1998), and it is a combination of the other dimensions described above. However, some datasets might appear of high quality with regards to all other dimensions, and be of great use to a certain problem, but might not be suited for other problems.

# 3 Methodology

This section will provide an overview of the research methodology, and the methods used to address the research questions and carry out the project.

Research methodology refers to the systematic and structured approach used to conduct research and gather information in order to address a specific research question or objective (Kothari 2004). It involves the overall strategy and techniques employed to collect, analyze, and interpret data, as well as the logical and theoretical framework guiding the research process (Karlsson 2010). The research in this thesis is based on empirical observations from a case study, and follows an inductive argumentation approach. The procedure of an inductive approach as described by Karlsson (2010) is illustrated in Figure 10, where observations are based on empirical evidence, results are generated through data analysis, and rules are derived from theory.

OBSERVATION
RESULT
RULE

Figure 10: The structure of a inductive research approach

In this research, the case study on TPL have been the foundation for observations, which have been used to derive results, before the implications of these results are discussed, which has led to new rules.

The research in this thesis is based on analyses of both qualitative and quantitative data. Qualitative data refers to descriptive data that is not possible to measure and needs to be expressed as text (McLeod 2019). Quantitative data refers to data that can be quantified and expressed as numbers (McLeod 2019). The following table categorizes the data used to derive the results of this thesis:

| **Sources of Qualitative Data** | **Sources of Quantitative Data** |
|---|---|
| - Interviews with staff at TPL | - Statistics from Bibliofil |
| - Meetings with staff at TPL | - Statistics from sorting logs |
| - On-site observations at TPL | - Time measurements of material handling processes at TPL |

Table 3: Categorization of qualitative and quantitative data

Qualitative research is considered to be great for deepening the understanding of a certain problem that cannot be quantified (Queirós et al. 2017), which in this case was understanding how material handling is performed at a library. The qualitative results are presented in a case study of the case company, TPL. Furthermore, the latter part of the case study involves quantitative results from data analysis of data collected through on-site observations. Results stemming from quantitative methods are considered to be objective, and representative of a larger population (Queirós et al. 2017). Furthermore, quantitative research is known to increase the validity of the findings from a qualitative study, and has been argued to propose a deeper and broader understanding of the studied phenomenon (Hurmerinta-Peltomäki and Nummela 2006). Lastly, the results obtained from utilizing predictive analytics to analyze and interpret data generated by the LMS, Bibliofil, are presented in Section 5.

## 3.1 Case Study

To acquire practical insights into library processes and establish a foundation for prediction models, a case study was conducted focusing on TPL. RQ1 and RQ2 require knowledge of a process specific to a library, and this knowledge was thought to be effectively obtained through practical experiences, rather than in literature. Eisenhardt (1989) emphasizes that theory derived from case study research often possesses strengths in terms of novelty. This makes it particularly suitable for exploring new research areas or areas where existing theories appear insufficient. In the context of material handling in public libraries and the potential enhancements brought by data analytics, the case study approach was deemed appropriate given the scarcity of existing theoretical frameworks, emphasized in Section 1.1.

The SmartLIB project was initiated in the spring of 2021. Prior to the specialization project carried out last semester, the only research conducted in the SmartLIB project was in the form of a master thesis by Illaria Caccese, which focused on book distribution between the different branches of the library. In the case study, Caccese's thesis has been utilized as a source of general information and statistics on Trondheim Public TPL. It served as a valuable secondary data source alongside other data and information provided by TPL. However, specific operational details such as book sorting, material flow, and material handling were largely unknown. Consequently, the foundation for conducting the case study was established by gathering comprehensive information on how the library conducts its operations and associated statistical data. This was accomplished through various methods, which are presented in the following sections.

### 3.1.1 Meetings, Interviews, and On-Site Observations

Meetings and interviews have been sources of qualitative data, conducted to obtain an understanding of library operations, particularly material handling tasks, and related challenges. Additionally, on-site observations were performed in order to attain a first-hand experience with the library processes and gather quantitative data. The combination of meetings, interviews, and on-site observations helped attain a holistic view of the library processes.

**Meetings**

Meetings have primarily been held with the IT department, and have provided general information about the library. Additionally, the meetings provided an opportunity to discuss technological visions, address challenges in library operations, particularly related to

IT solutions, and gain insights into the library's LMS, Bibliofil, including data extraction methods. The library staff demonstrated a strong willingness to embrace digitization and automation in library processes, displaying enthusiasm for future technological implementations. This positive attitude towards change management is considered a crucial success factor, as outlined in Table 1 in Klein and Sorra (1996).

The meetings followed a semi-structured interview approach, combining predefined questions with open discussion. The predefined questions ensured that key areas of interest were addressed, while the general discussion allowed for the exploration of additional knowledge and insights that may not have been covered by the predefined questions. This approach is consistent with the findings of Barriball and While (1994), who concluded that semi-structured interviews can elicit complete information and clarify interesting and relevant.

### Interviews

Once a general understanding of the daily operations and associated challenges was obtained, it was deemed interesting to consult with the employees who have on-hand experience with the different operations in the library. These individuals possess firsthand knowledge of the specific issues explored in this thesis, making them valuable sources of expert domain knowledge. Therefore, interviews were conducted with librarians responsible for various sections of the library, including the sorting area, the children and youth area, and the second floor.

These interviews provided in-depth information about the material handling tasks present in the library, particularly those related to managing the incoming book flow. Furthermore, time spent on manual material handling tasks was discussed with these librarians, in order to cross-validate the recorded times stemming from on-site observations, presented in the next paragraph.

### On-Site Observations

The case study heavily relied on on-site observations to generate findings, as no existing data was available regarding how librarians allocate their time to various tasks during their shifts. Therefore, primary data was gathered through on-site observations, specifically to measure and document the time dedicated to material handling tasks and the duration of each individual task. The durations were manually recorded using stopwatches. Examples of observed tasks include activities such as transferring full bins from the sorting machine to temporary shelves, unloading incoming books from other branches into the sorting machine, and managing non-collected reservations. These tasks are further described in Section 4. This data collection process was conducted in collaboration with both the library staff and a research team from Georgia Institute of Technology, who were also involved in the SmartLIB project.

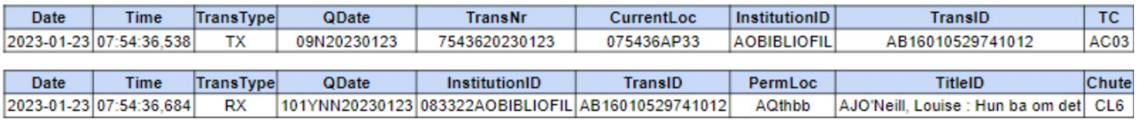
### 3.1.2 Data Analysis and Performance Simulation

Lastly, existing data stemming from the sorting machine, have been used to attain results in the case study. The sorting logs have enabled the possibility to analyze the book distribution in the sorting machine for a large period of time, ensuring that the results derived were based on large sample sizes which increased the robustness of the results. Based on this, it was possible to propose alterations to the sorting setup.

In order to investigate how different sorting setups affected time spent on material handling, performance simulations were performed. The simulation used historical data to determine incoming book flow in combination with the time measurements related to the different material handling tasks, gathered through observations

#### Dataset Explanation - Sorting Logs

For every book passing through the machine, two types of transactions are generated; TX and RX. TX transactions are queries, while RX transactions are responses. The self-return machines produce separate log files every single day. Therefore, it is necessary to combine the files to obtain a dataset describing the sorting for a longer period of time.



| Date | Time | TransType | QDate | TransNr | CurrentLoc | InstitutionID | TransID | TC |
|---|---|---|---|---|---|---|---|---|
| 2023-01-23 | 07:54:36,538 | TX | 09N20230123 | 7543620230123 | 075436AP33 | AOBIBLIOFIL | AB16010529741012 | AC03 |

| Date | Time | TransType | QDate | InstitutionID | TransID | PermLoc | TitleID | Chute |
|---|---|---|---|---|---|---|---|---|
| 2023-01-23 | 07:54:36,684 | RX | 101YNN20230123 | 083322AOBIBLIOFIL | AB16010529741012 | AQthbb | AJO'Neill, Louise : Hun ba om det | CL6 |

Figure 11: Overview of the format of the sorting logs

Figure 11 illustrates the format of the TX and RX transactions. The codes present in the last three fields of the TX transactions are explained as follows:

- AO = institution id.
- AB = item identifier.
- AC = terminal code.

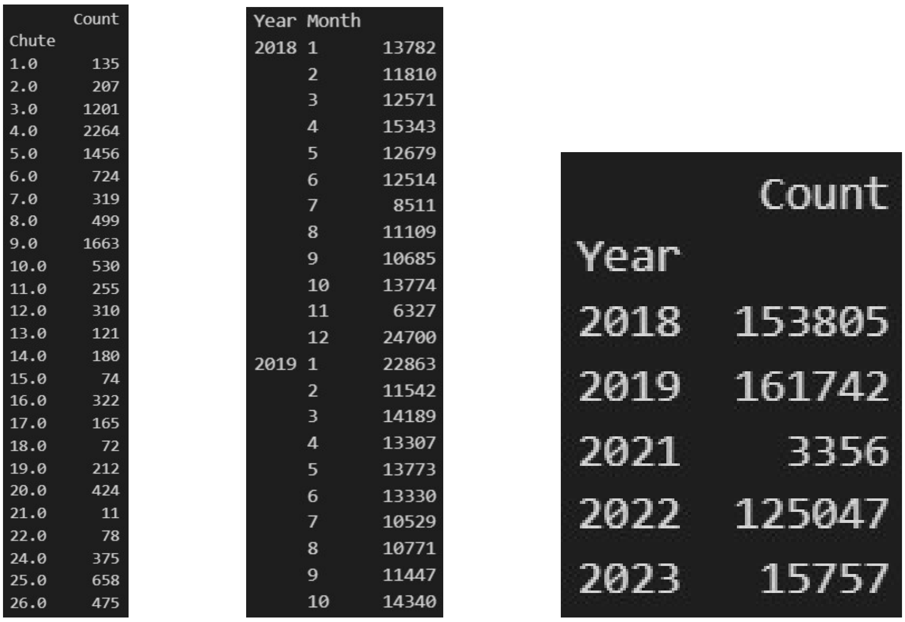
The AO and AB codes are also present in the RX transactions. These are used to link the query and response. The RX transactions also contain other codes:

- AQ = permanent location.
- AJ = title identifier.
- CL = sort bin.

The AQ code describes the owner branch of the book. This code provides information about where the book is to be transported after being sorted in the sorting machine. The

AJ code holds information about the author of the book in addition to the book title. Lastly, the CL code describes which bin the book should be sorted into.

As mentioned, the three return machines produce their own sorting logs for each day. To enhance data comprehensibility, a script was utilized to merge these individual files. During the merging process, particular attention was given to the RX transactions, as they contain information such as the book's permanent location, title, and sorting details. Based on these parameters, the sorting logs were capable of generating statistics regarding the daily, monthly, and yearly book throughput in the sorting machine, as well as the distribution of books across different bins. All of the TX transactions were disregarded and excluded from the dataset.



```
       Count
Chute
1.0      135
2.0      207
3.0     1201
4.0     2264
5.0     1456
6.0      724
7.0      319
8.0      499
9.0     1663
10.0     530
11.0     255
12.0     310
13.0     121
14.0     180
15.0      74
16.0     322
17.0     165
18.0      72
19.0     212
20.0     424
21.0      11
22.0      78
24.0     375
25.0     658
26.0     475
```

(a) Distribution of books in bins for one day

```
Year Month
2018 1      13782
     2      11810
     3      12571
     4      15343
     5      12679
     6      12514
     7       8511
     8      11109
     9      10685
     10     13774
     11      6327
     12     24700
2019 1      22863
     2      11542
     3      14189
     4      13307
     5      13773
     6      13330
     7      10529
     8      10771
     9      11447
     10     14340
```

(b) Number of books delivered in the staff sorting machine by month

```
       Count
Year
2018  153805
2019  161742
2021    3356
2022  125047
2023   15757
```

(c) Number of books delivered in the staff sorting machine by year

Figure 12: Examples on sorting logs statistics

### 3.1.3 Summary of Case Study Methods

The figure below summarizes the case study methods and the main outcome of each method:

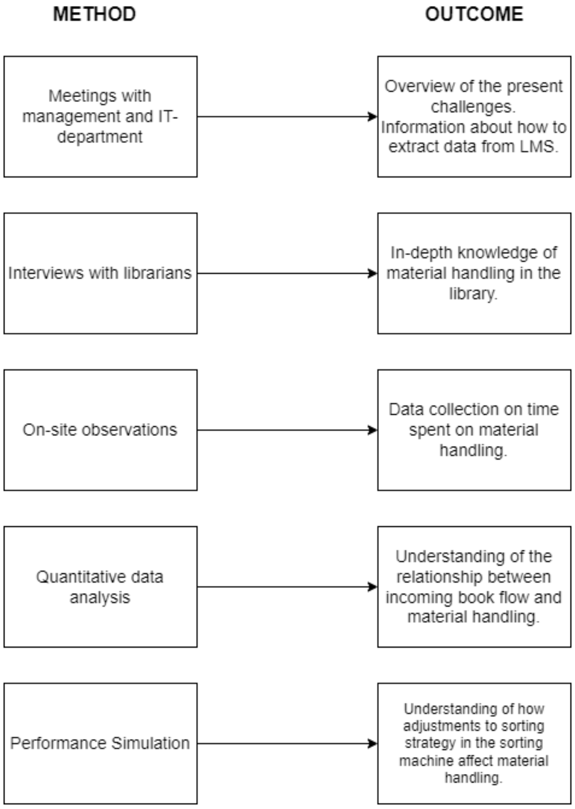

METHOD
OUTCOME
Meetings with management and IT-department
Overview of the present challenges. Information about how to extract data from LMS.
Interviews with librarians
In-depth knowledge of material handling in the library.
On-site observations
Data collection on time spent on material handling.
Quantitative data analysis
Understanding of the relationship between incoming book flow and material handling.
Performance Simulation
Understanding of how adjustments to sorting strategy in the sorting machine affect material handling.

Figure 13: Methods used to conduct the case study and their main outcome.

## 3.2 Predictive Modeling

In addition to the sorting logs, TPL is in possession of large quantities of data, regarding the number of visitors, loan transactions, and information about each individual book. All of this data is stored in the LMS, and provide information about historical demand. In order to further examine the applicability and feasibility of predictive analytics to predict incoming book flow in a library, predictive models based on machine learning, were developed and applied to predict loan lengths. In theory, predictive modeling might seem beneficial for the areas explored in this thesis. The machine learning modeling helped verify the effects.

The conceptual framework illustrated in Figure 14 was derived by analyzing 48 papers, which dealt with machine learning for time estimation problems, in a systematic literature review conducted last semester in the specialization project. It is a fairly standard approach for the development of a machine learning model, and has been followed in this research.

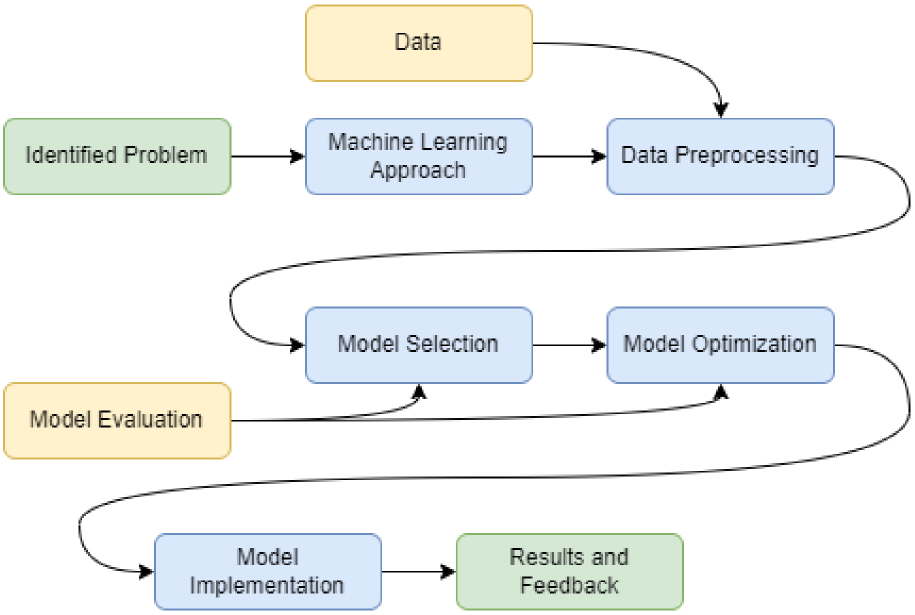
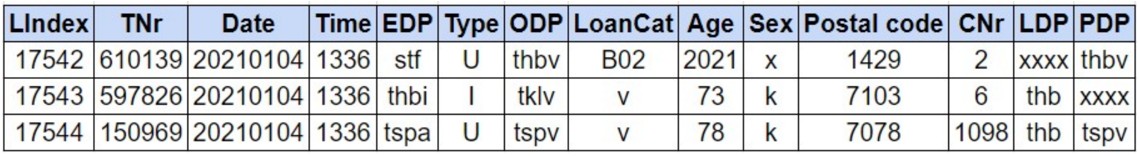


Figure 14: Conceptual framework for the application of machine learning for time estimation.

Firstly, a problem has to be identified. In this particular case, the problem was identified through the work conducted in the specialization project, and it was further refined based on the results obtained from the case study. Consequently, the problem to be addressed using machine learning techniques is the prediction of the incoming book flow to the main library by predicting the loan durations. The desired output in this problem is a continuous variable, specifically the number of days representing the loan duration. Therefore, the machine learning approach employed can be characterized as regression. This chapter will outline the process of adapting the initial dataset and provide a comprehensive explanation of the various steps involved in developing the machine learning model, as well as an overview of the dataset itself.

### 3.2.1 Dataset - Loan and Return Statistics Generated from Bibliofil

In order for a machine learning model to operate effectively, it relies on input data, often in significant volumes. As previously mentioned, TPL generates a continuous stream of data. For instance, TPL possesses substantial quantities of data related to loan and return transactions, which are stored in their LMS, Bibliofil. This section aims to provide an explanation of the dataset containing loan and return transactions. The dataset encompasses all transactions that occurred in the years 2021 and 2022, organized in the format illustrated in Figure 15.



| LIndex | TNr | Date | Time | EDP | Type | ODP | LoanCat | Age | Sex | Postal code | CNr | LDP | PDP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17542 | 610139 | 20210104 | 1336 | stf | U | thbv | B02 | 2021 | x | 1429 | 2 | xxxx | thbv |
| 17543 | 597826 | 20210104 | 1336 | thbi | I | tklv | v | 73 | k | 7103 | 6 | thb | xxxx |
| 17544 | 150969 | 20210104 | 1336 | tspa | U | tspv | v | 78 | k | 7078 | 1098 | thb | tspv |

Figure 15: Overview of the raw transaction data

**Column 1 ("LIndex"):** Is the index of the transaction in the dataset.

**Column 2 ("TNr"):** Title number of the book. Exclusive for each title, but not exclusive for the copy of each title.

**Column 3 and 4 ("Date" and "Time"):** Is the date- and timestamp associated with the transaction, in the formats YYYYMMDD and HHMM.

**Column 5 ("EDP"):** Executive department. Describes where the transaction was executed, with branch and department.

**Column 6 ("Type"):** Type of transaction. The transaction types relevant to this thesis are the ones beginning with "I" and "U", which are returns and loans, respectively.

**Column 7 ("ODP"):** Owner department. As of now, all books are returned to their owner department if it is returned to a different branch. Uses the same acronyms as the fourth column.

**Column 8-11 ("LoanCat", "Age", "Sex" and "Postal code"):** Descriptive information about the loaner. Loan category is divided into three categories: children (under 15), adults, and institutions (schools, kindergartens, other libraries, etc).

**Column 12 ("CNr"):** Copy number. This number functions as an ID for the copy in combination with the title number.

**Column 13 and 14 ("LDP" and "PDP"):** LDP refers to the loaner's "home" department. No one in the library staff, or from Bibliofil, has been able to provide an explanation of PDP, and it has therefore been neglected in the dataset.

The acronyms used to refer to the different branches, used in EDP, ODP, LDP, and PDP are explained in the table below:

| **Library Branch Acronym** | **Library Name** |
|---|---|
| thb | Trondheim folkebibliotek Hovedbiblioteket |
| tby | Trondheim folkebibliotek Byåsen |
| tra | Trondheim folkebibliotek Ranheim |
| trv | Trondheim folkebibliotek Risvollan |
| tkl | Trondheim folkebibliotek Klæbu |
| tmo | Trondheim folkebibliotek Moholt |
| tsp | Trondheim folkebibliotek Saupstad |
| the | Trondheim folkebibliotek Heimdal |
| tbu | Trondheim folkebibliotek Buran |
| tfe | Trondheim folkebibliotek Fengsel |
| mpmi | MappaMi |

Table 4: Overview of the acronyms for the different branches

Some of the acronyms come with an additional fourth descriptive letter, for example, b for "barn" (children), u for "ungdom" (youths) and v for "voksen" (adult), which describes which area of the library the transaction is related to. For the main library, with the abbreviation “thb”, the fourth descriptive letter can also indicate which counter was used for the return transaction. The code “mpmi” refers to “MappaMi”, which is the web portal of the library, where patrons can handle their loans and reservations. This only appears as a value for EDP.

### Initial Preparation of the Dataset

To be able to further analyze the dataset, the dataset had to be in a compliant format. Therefore, the transaction data was exported in .csv format, before it was converted into a more manageable format, namely DataFrames from the Pandas library. Pandas DataFrame is one of the most used and well-documented tools for data analytics and machine learning in Python. After the data was converted to DataFrames the cleaning process was initialized.

Since the dataset originally included all of the transactions performed in the library system, transactions that were deemed irrelevant for the purpose of this thesis were removed. Only the transactions that were variants of type “U” and “I” were included. Furthermore, all “I” transactions with another executive department than the main branch, were removed so that only the incoming book flow at the main branch was considered. “U” transactions with EDP equal to “mpmi” were removed, as these transactions actually do not describe a loan, but rather a loan extension.

Moreover, for the dataset to provide information about loan lengths, the loan transactions had to be linked to their corresponding return transaction. Unfortunately, there exists no ID that links a loan- and return transaction to each other. However, because of the chronological ordering of the transactions, it was possible to link the loan and return transactions. This was achieved by matching the title- and copy number of the transactions while ensuring that the transaction number related to the return transaction was greater than the loan transaction. It was important to consider that the number of “I” transactions far exceeds the number of “U” transactions in the dataset. This can be explained by “I” transactions being created for several activities other than the initial return, such as when a book is prepared for reservation and when a book is moved from one branch to another. However, when consecutive “I” transactions are created for the same book, they are typically characterized with almost equal timestamps, thus not affecting the loan length. It is the first registered “I” transaction that appears after a “U” transaction that has been considered in these cases. In total, in excess of 470 000 transactions were remaining after the dataset had been prepared.

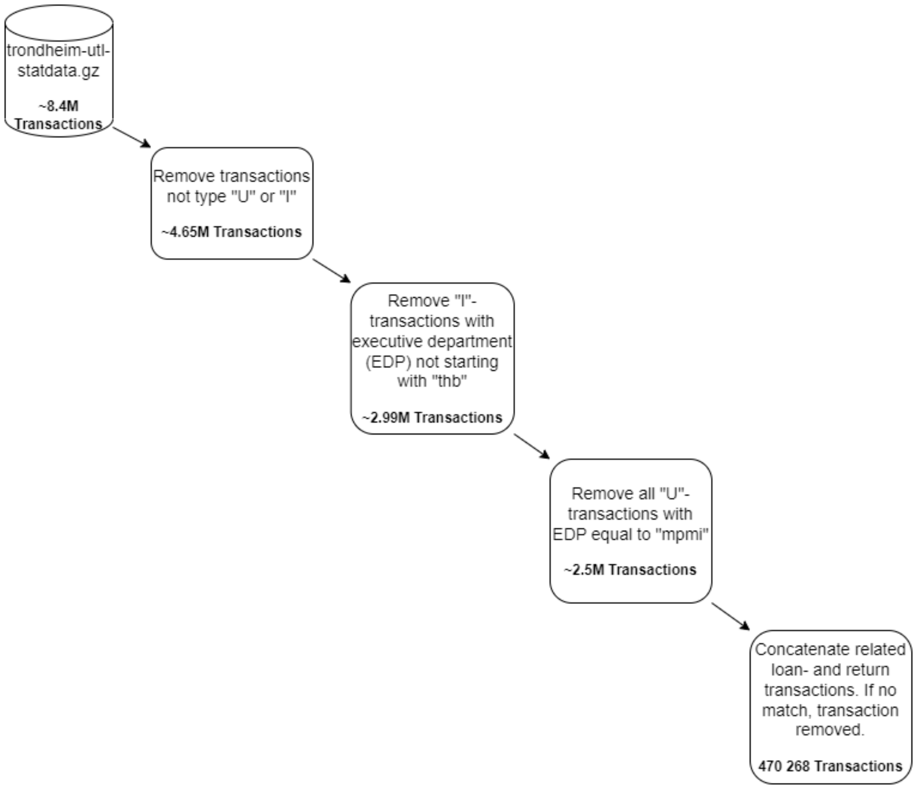

trondheim-utl-statdata.gz
~8.4M Transactions
Remove transactions not type "U" or "I"
~4.65M Transactions
Remove "I"-transactions with executive department (EDP) not starting with "thb"
~2.99M Transactions
Remove all "U"-transactions with EDP equal to "mpmi"
~2.5M Transactions
Concatenate related loan- and return transactions. If no match, transaction removed.
470 268 Transactions

Figure 16: Dataset filtering process.

### 3.2.2 Data Preprocessing

As mentioned in Section 2, data preprocessing is an important step when building machine learning models. The term “garbage in, garbage out” is commonly used for machine learning applications, indicating that a model’s performance is restricted by the quality of the input data. The preprocessing was separated into four parts:

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction

For this case, data integration was not relevant since the input data only has one data source, Bibliofil, and all data was in the same dataset. All the preprocessing steps were carried out in Python using Pandas Dataframes as data format. Furthermore, the Matplotlib library was used for model visualization and analyzing features.

Before the initial step of preprocessing, data cleaning, the variable object types were checked. The variables with information about the date and time were converted to the datetime and timedelta objects. This can be described as mandatory data transformation, for the objects to be readable for a machine learning model. Furthermore, after concatenating the loan- and return transactions, it became apparent that there existed more department codes than what was described by Bibliofil, such as a book bus traveling around the Trondheim region. Therefore all transactions with deliveries to other branches than the main library ("thb") or the county library ("stf") that were generated in the concatenation process were also removed from the dataset. This constituted approximately 5 000 transactions.

### Data Cleaning

To increase the quality of the dataset, cleaning was performed after the filtering process. Cleaning has mainly consisted of handling missing values and dealing with outliers and noise in the dataset, explained in the paragraphs below.

**Missing Values and Faulty Data:** The initial stage of data cleaning involved examining the dataset for erroneous data entries and missing values, as the presence of such instances can distort results and diminish the reliability of data analysis. To accomplish this, the distributions of values for the various variables were checked. These assessments revealed that all rows in the dataset were complete, without any missing values. However, a sporadic occurrence of values such as "ukjent", "u", or "xxxx" was observed, indicating unknown or unregistered information. These values are automatically generated by Bibliofil when there is missing information in a transaction, such as the unavailability of the sex or postal code of the individual executing the transaction.

To address unknown values for the loaner's department, corrections were made based on analyzing the values available in the dataset for this variable. These values were assigned the same department as the loan executive for the corresponding loan. A similar approach was employed for the owner department. This correction process affected 80,766 transactions for the loaner's department and 8,716 transactions for the owner department.

In contrast, unknown values for gender were not assigned new values due to their significant representation within the overall transaction dataset, illustraed in Figure 17. Accurately assigning a new value in such cases proved challenging. Furthermore, gender was not thought to have a noteworthy impact on the target variable, which was confirmed later in the preprocessing.

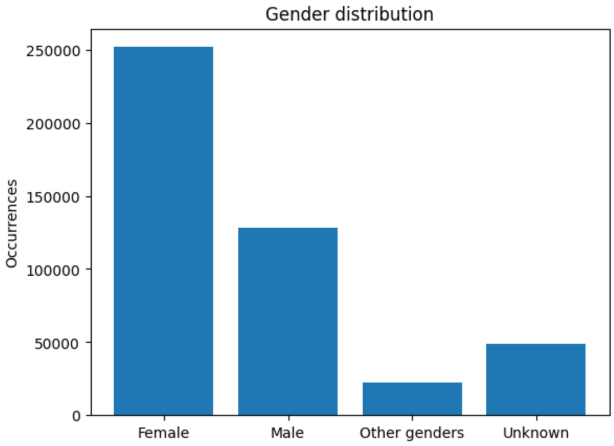


Figure 17: The gender distribution in the dataset, where a large portion of "unknown" values is present.

Inspection of the dataset also uncovered errors present in the dataset related to the age variable. It was observed that some age values far exceeded what can be believed to be possible, while some transactions had negative values for age. These values were handled with simple imputation, and new values were assigned and normally distributed based on the age distribution of the other transactions.

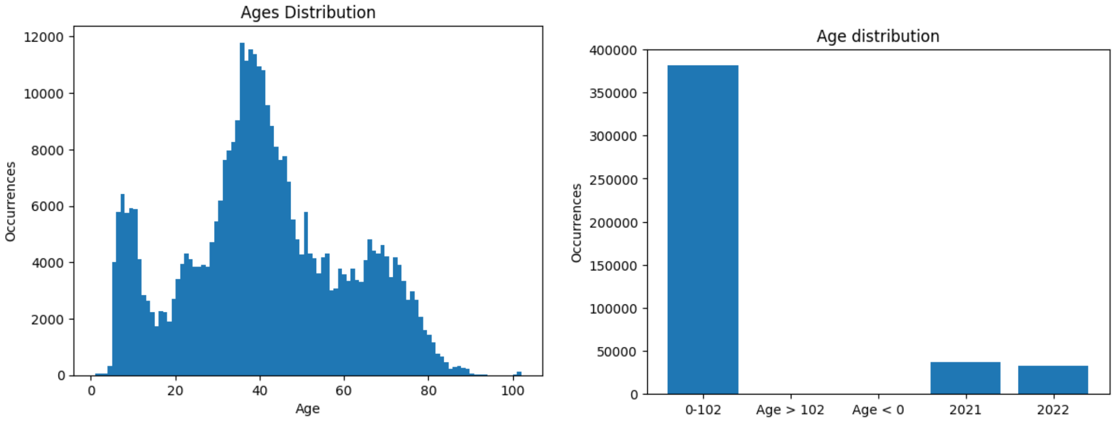
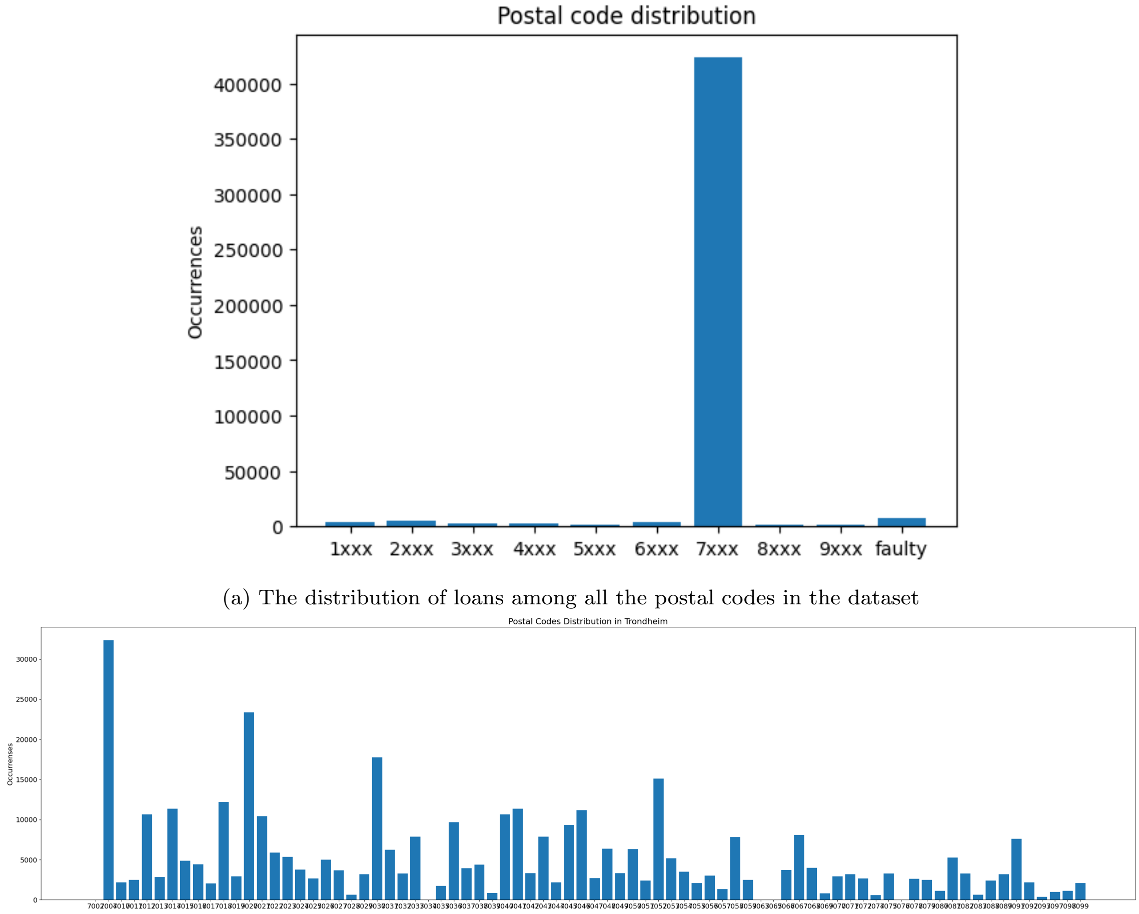


Figure 18: Age distribution and occurrences of extreme age values in the dataset.

Approximately 95% of the transactions in the dataset were associated with postal codes beginning with "7XXX". This observation can be attributed to the fact that postal codes within the municipality of Trondheim typically commence with "70XX". However, there were instances of postal codes in the dataset that did not start with "7", indicating patrons from other municipalities who requested books available only at TPL.

Nevertheless, around 7,000 transactions in the dataset contained either non-existing or invalid postal codes. To address this issue, a replacement procedure was implemented, utilizing a probability distribution based on the presence of valid postal codes within the municipality of Trondheim as observed in the dataset. The specific distribution is illustrated in Figure 19b, providing insights into the probability distribution used for replacing non-existing or invalid postal codes.



(a) The distribution of loans among all the postal codes in the dataset

(b) Distribution of loans among the postal codes within the municipality of Trondheim

Figure 19: The distribution of postal codes in the transactions.

**Outliers and Noise:** In order to ensure the accuracy and reliability of the machine learning models, measures were taken to handle outliers and noise in the dataset. One particular aspect addressed was transactions with a loan period of zero days, which were considered as noise and subsequently removed from the dataset. These instances typically corresponded to loans associated with library events and did not reflect the regular loaning patterns of patrons. In total, this constituted 14,084 transactions that were excluded.

On the other hand, loan durations exceeding 50 days were not identified as outliers, as they were more likely the result of loan extensions rather than delayed returns. Consequently, these transactions were not treated as outliers but were instead recognized as contributing to the natural variation and diversity within the dataset. This behavior represents a normal aspect of the data distribution rather than an anomaly.

### Data Transformation

Data transformation was performed to both ensure data compatibility and increase the quality of the dataset. Firstly, the mandatory transformation of converting non-numeric features into numeric features was performed. Considering the aim of the prediction was to estimate a loan length based on two dates, a variable containing the difference between the loan- and return date was created, and used as the target variable.

The dataset contained several features that had non-numeric information, necessitating their conversion. These features included the executive department, owner department, loan category, sex, and loaner department. The loan category feature consisted of 66 unique variables, which were grouped into 8 variables to simplify the representation based on their characteristics. For example, variables like “B04” and “B16” were categorized as subgroups of the “b” value, which denoted children. The subsequent digits describe age, which was considered redundant information, as it was already described in the age variable as well.

Various techniques can be employed to convert non-numeric variables into numeric representations. In this study, dummy values and indexed features were utilized. The creation of dummy values involved assigning a separate feature for each unique value in the original feature. This technique is most suitable for features with a small number of distinct values and was applied to the loan category, sex, and book category features. On the other hand, the indexed feature technique was employed for the executive, owner, and loaner departments. These features shared the same abbreviation for different departments. The conversion process is described in detail in Table 5. Additionally, a category labeled as “others” was created to group libraries located outside the municipality of Trondheim, which occurred in only a few transactions.

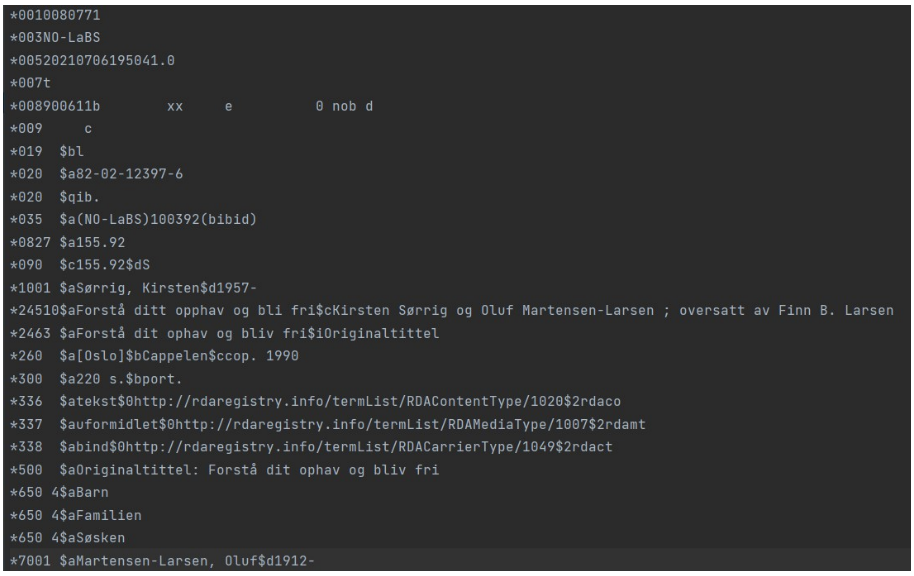
| Branch | Index | Branch | Index |
|---|---|---|---|
| thb | 1 | tsp | 7 |
| tby | 2 | the | 8 |
| tra | 3 | tbu | 9 |
| trv | 4 | tfe | 10 |
| tkl | 5 | stf | 11 |
| tmo | 6 | others | 12 |

Table 5: Index feature conversion for the features: EDP, ODP and LDP

Furthermore, the date, time, and loan length features needed to be converted from datetime and timedelta objects into integer objects. In the conversion process, the date feature was converted into three new integer features describing year, month, and day. The time feature was converted into new integer features describing hour and minute. The loan length feature was also converted into an integer feature.

It was decided not to perform any feature normalization since most features were categorical and could not be compared on a similar scale. While “Age” could have been a feature suitable for normalization, normalizing only one variable would not provide significant benefits and would therefore only contribute to decreasing the interpretability of the dataset. Additionally, since most variables were already categorical, there was no need to perform any feature discretization, since the remaining numerical values could lose information if discretized.

Feature engineering is a widely employed technique aimed at improving the performance of machine learning models. It involves leveraging domain knowledge to create new predictor variables that provide valuable insights within the dataset. Given the limited information available about each loaned book in the dataset, the potential for feature engineering was explored to incorporate additional descriptive information about the books. The easiest way of achieving this would have been to merge the transaction dataset with a dataset containing descriptive information about each title number, such as the book catalog dataset. However, the book catalog dataset was found to be in a distinct format, with a considerable number of missing values and redundant fields, posing challenges in its interpretation and utilization. Consequently, it was deemed too tedious to engineer features describing additional book information.



```
*0010080771
*003NO-LaBS
*00520210706195041.0
*007t
*008900611b        xx     e          0 nob d
*009     c
*019  $bl
*020  $a82-02-12397-6
*020  $qib.
*035  $a(NO-LaBS)100392(bibid)
*0827 $a155.92
*090  $c155.92$dS
*1001 $aSørrig, Kirsten$d1957-
*24510$aForstå ditt opphav og bli fri$cKirsten Sørrig og Oluf Martensen-Larsen ; oversatt av Finn B. Larsen
*2463 $aForstå dit ophav og bliv fri$iOriginaltittel
*260  $a[Oslo]$bCappelen$ccop. 1990
*300  $a220 s.$bport.
*336  $atekst$0http://rdaregistry.info/termList/RDAContentType/1020$2rdaco
*337  $auformidlet$0http://rdaregistry.info/termList/RDAMediaType/1007$2rdamt
*338  $abind$0http://rdaregistry.info/termList/RDACarrierType/1049$2rdact
*500  $aOriginaltittel: Forstå dit ophav og bliv fri
*650 4$aBarn
*650 4$aFamilien
*650 4$aSøsken
*7001 $aMartensen-Larsen, Oluf$d1912-
```

Figure 20: Record from the book catalog dataset.

Instead, it was focused on deriving new features from the features available in the transaction dataset. For instance, a feature describing the book category was created based on the value of the owner department feature. As explained previously, the first three characters of the value in this feature indicate the owner branch of the book, while the fourth character describes the specific section within the branch to which the book is assigned. Examples include “v”, adult section, and “b”, children and youth section. This fourth character was used to create the “Book Category” feature.

Moreover, a new feature was introduced to capture the specific day of the week on which the loan and return transactions occurred. This additional feature aimed to enable the models to better identify and incorporate the weekly variations in loan and return patterns. It was created by using a built-in function within the date library in Python. The different weekdays were converted into integers from one to seven, to represent the weekdays as numerical values.

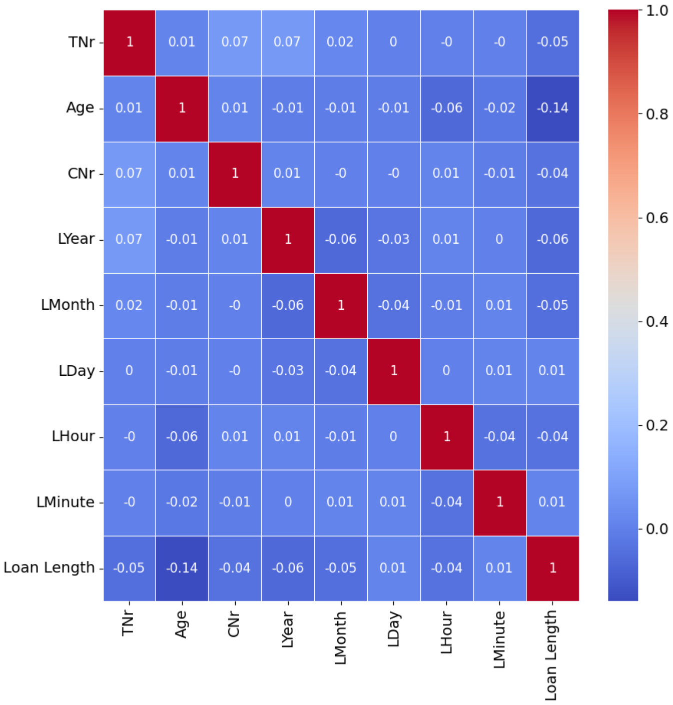
However, given the limitations of the available data and the observation that the introduction of new features did not significantly enhance the predictive performance of the models, it was determined that dedicating extensive time to creating additional features would not be warranted.

### Data Reduction

Finally, to eliminate redundant features and reduce computational complexity, a process of data reduction was undertaken. Upon concatenating the loan and return transactions, it was determined that the only useful information from the return transaction was the date feature, which was utilized to calculate the target value. The remaining features associated with the return transaction were either deemed inconsequential for the prediction models, or they were already captured and described in the corresponding loan transaction. Hence, including them in the modeling process would introduce redundancy and offer limited additional value.

Additionally, a thorough examination of the concatenated transactions was conducted to assess the relevance of other features with respect to the prediction target variable, loan length. The feature labeled “LIndex” was deemed unnecessary for the dataset as its sole purpose was to serve as a validation tool during the mapping of loan and delivery transactions, offering no substantive information related to the target variable. Similarly, the type feature, which denoted whether a transaction was a loan or return, was deemed to be of little value after concatenation, as all records had identical values for this feature. As presented in Guyon and Elisseeff (2003), a feature with zero variability provides no additional information and can be considered redundant. Furthermore, after analyzing the data, it was determined that the PDP feature could be omitted. As mentioned, no one in the library or from Bibliofil could provide an explanation of this variable, “actual ownership department”, and it was observed to not differ from the ODP, unless it was unknown.

To perform additional feature selection, a correlation analysis was carried out to examine the relationships between variables in the dataset. However, no significant correlations were observed between any of the variables and the target variable. Therefore, it was not possible to establish a threshold for determining which variables should be retained or eliminated. Consequently, it was determined that no further variables would be removed from the dataset. The correlation matrix, illustrating the correlations among the variables, is depicted in the figure below.



| | TNr | Age | CNr | LYear | LMonth | LDay | LHour | LMinute | Loan Length |
|---|---|---|---|---|---|---|---|---|---|
| TNr | 1 | 0.01 | 0.07 | 0.07 | 0.02 | 0 | -0 | -0 | -0.05 |
| Age | 0.01 | 1 | 0.01 | -0.01 | -0.01 | -0.01 | -0.06 | -0.02 | -0.14 |
| CNr | 0.07 | 0.01 | 1 | 0.01 | -0 | -0 | 0.01 | -0.01 | -0.04 |
| LYear | 0.07 | -0.01 | 0.01 | 1 | -0.06 | -0.03 | 0.01 | 0 | -0.06 |
| LMonth | 0.02 | -0.01 | -0 | -0.06 | 1 | -0.04 | -0.01 | 0.01 | -0.05 |
| LDay | 0 | -0.01 | -0 | -0.03 | -0.04 | 1 | 0 | 0.01 | 0.01 |
| LHour | -0 | -0.06 | 0.01 | 0.01 | -0.01 | 0 | 1 | -0.04 | -0.04 |
| LMinute | -0 | -0.02 | -0.01 | 0 | 0.01 | 0.01 | -0.04 | 1 | 0.01 |
| Loan Length | -0.05 | -0.14 | -0.04 | -0.06 | -0.05 | 0.01 | -0.04 | 0.01 | 1 |

Figure 21: Correlation analysis chart of the numerical values in the dataset.

### 3.2.3 Model Selection

During the development of various models, the Scikit-learn library was employed. Scikit-learn is a Python machine learning library that provides a comprehensive range of supervised and unsupervised learning models, performance metrics, and auxiliary functions for building machine learning models. An advantage of Scikit-learn is its seamless integration with Pandas Dataframes, which were utilized throughout the preprocessing phase. This integration facilitated efficient data manipulation and preprocessing tasks.

#### Machine Learning Algorithms

The selection of machine learning models for the prediction problem was based on the findings of a systematic literature review conducted during the previous semester's specialization project. The results of the literature review highlighted that various machine learning algorithms can effectively predict time estimations, and emphasized the importance of comparing multiple models, as the optimal model choice depends on the specific case. Among the models frequently mentioned in the reviewed literature Artificial Neural Network, Decision Tree, and Support Vector Regression consistently demonstrated satisfactory performance across different problem domains. Consequently, these three models were initially employed to predict the incoming book flow. However, it was observed that the DT overfitted. To address this issue, ensemble models such as Random Forest, Adaptive Boosting, and Gradient Boosted Decision Trees were also incorporated in the analysis.

#### Hyperparameter Tuning

To determine the optimal combination of hyperparameters for the machine learning models, a grid search algorithm was employed during the model fitting process. Hyperparameters are parameters that define the behavior and configuration of the models, and their composition significantly impacts the model's performance. As there is no predetermined best combination of hyperparameters, the grid search algorithm systematically explores various combinations by running the machine learning algorithm multiple times with different hyperparameter settings. The algorithm then evaluates the performance of each combination and selects the model with the hyperparameter composition that yields the best results.

#### Train-test Split

To ensure the validity of the machine learning models' performance, the preprocessed dataset was divided into separate training and testing sets. Training and testing the models on the same data can lead to overfitting, where the models perform well on known data but fail to generalize to unseen data. To mitigate this issue, the dataset was randomly split into a training set and a testing set using a 70:30 ratio. This allocation ensures that 70% of the data is used for training the models, while the remaining 30% is reserved for evaluating their performance on unseen data. By randomly distributing the data between the train and test sets, it is ensured that both sets are representative of the original dataset, capturing the variability and patterns present in the data. This approach allows for a robust assessment of the models' predictive capabilities on unseen data and helps to prevent overfitting.

### 3.2.4 Model Evaluation

During model selection, when comparing several models, the different models were evaluated in order to select the best performing. Firstly, their performance was evaluated, before the quality of the input data was assessed.

#### Performance Evaluation

The performance of the models was evaluated based on different evaluation metrics. To attain a holistic view of the models' performance, multiple metrics were used. Botchkarev (2018) and Chai and Draxler (2014) explains: "as every statistical measure condenses a large number of data into a single value, it only provides one projection of the model errors emphasizing a certain aspect of the error characteristics of the model performance". Therefore, using several evaluation metrics reduces the likelihood of wrongly evaluating a model. Additionally, choosing only one evaluation metric can be challenging, as there exists no universally best evaluation metric (Silver et al. 1998). It was decided to use MAE, MAPE, RMSE, and $R^2$ for evaluating the models, as they were the most frequently used metrics for regression problems in the papers analyzed in the systematic literature review, described earlier. When evaluating the performance of the models, the performance of the models on both the train and test sets was considered.

### Data Quality Assessment

To further understand the poor results provided by the machine learning models, it was deemed desirable to assess the quality of the input data. The transactional dataset was analyzed, and the data quality was assessed based on the dimensions presented in Section 2.4. Fitness for purpose was difficult to measure and was decided based on experience and domain knowledge.

# 4 Incoming Book Flow and Related Material Handling at Trondheim Public Library

The main purpose of the SmartLIB project was previously referred to as "developing a smart library system that will solve current limitations". One of these limitations is that too much time is spent on material handling tasks each day. This thesis focuses on investigating the benefits yielded from utilizing predictive analytics in a public library, by predicting incoming book flow, an unexplored research area. However, before applying predictive analytics with the aim of improving processes in the library, it is imperative to first be familiar with the library operations and identify challenges in the current system. Consequently, a case study has been conducted to attain knowledge regarding library operations. This case study provides insight into how the incoming material flow affects and initializes material handling at the main library, which is the foundation for answering the first and second RQs.

In this chapter, general information about TPL is presented, before the general notions of incoming material flow are described. After the material handling tasks necessary to handle the incoming material flow have been presented, results from the data collection regarding the amount of time library employees devote to different material handling tasks are presented. Lastly, the various challenges that were found related to the sorting machine are presented and examined in greater detail, before the effects of alternative sorting setups are explored.

## 4.1 Trondheim Public Library - Main Branch

The main library is located in the city center and experiences the most activity. The main library spans four floors and a basement, with the first, second, and third being open to patrons. The main storage is located in the basement, alongside a garage. The main library is the only branch equipped with an automated sorting process of incoming books, performed by a book sorting machine located on the first floor. Moreover, the first floor is dedicated to children and is frequently visited by kindergartens and primary schools. The first floor also houses comics and temporary exhibitions. Fiction can be found on the second floor, while non-fiction is spread between the second and third floors. The second and third floors do also have multiple seating areas and reading spaces available, attracting both elderly individuals and students who utilize the library throughout the day, to either read newspapers or study. The library is not only a place where patrons collect and return books, but also a gathering spot for the public.

### 4.1.1 Return Statistics

The main library accounts for approximately one-third of all the submissions at Trondheim Public Library. The returns to the main library consist of both returns directly by patrons as well as returns of books coming with the truck from other branches. A trend over the last years shows that the number of returns to the main library has decreased by 33% from its peak, reaching a historical low during the Covid pandemic. Numbers from 2022 indicate that the numbers are recovering slowly, closer to the levels before the pandemic. The gradual decrease observed, excluding the years affected by the pandemic, can likely be contributed to the modernization and expansion of the branches at Moholt, Risvollan, and Saupstad, and the opening of new branches at Ranheim (2016), Klæbu (2020), and Buran (2022). This has likely increased the number of returns performed at other branches. Furthermore, the increasing popularity of e-books can also be influential. However, the library suggests that e-books only account for approximately 5% of total loans.

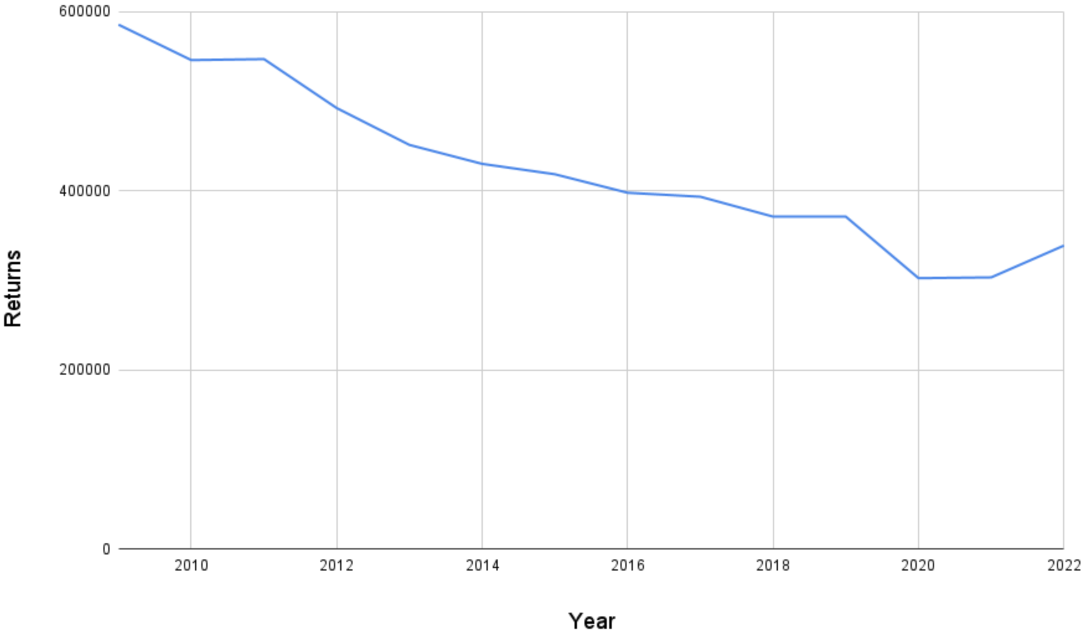


Figure 22: Number of returns each year at the main branch.

However, the total number of returns to the main library depicted in the graph might be somewhat imprecise, as it is based on books and media passing through the sorting machine, which omits returns being performed over the information counters at the library. For instance, books that are used in language courses are delivered to librarians at the counters.

### 4.1.2 Layout and Storage Policies

All libraries are dependent on operating with functioning storage systems, often with different policies for different parts of the library. Storage of books on the final shelves, where books are displayed, will differ from how the books not on display are stored. At TPL, the books that are not on display are stored in the basement area, described to be operating with a "chaotic" storage system. The "chaotic" storage system is their own version of what was described as random storage by Bakkali et al. (2013). In the rest of the library, all books have a designated place, thus operating with a dedicated storage policy.

It is possible to categorize the main floors of the library into two departments; a department for children and youth and a department for adults. The children and youth department is located on the first floor. Within the children and youth department, the books are placed based on languages, genres, and the level of reading, and then further placed based on the alphabetical order of authors.

The adult department spreads across all three floors of the library. The first floor holds exhibition books, that are regularly rotated, and comics. The books used for the temporary exhibition on the first floor are based on trends or events held at the library or in Trondheim. Furthermore, the second floor holds both fiction and non-fiction. The position of fiction books is determined by genre, and then further on the alphabetical order of the authors within the genre, while the placement of the non-fiction books is based on Dewey numbers. Dewey numbers is a classification system that allows for new books to be added to a library in their appropriate location based on the subject (Dewey 1876). The first digit represents the superior subject category or broader topic of the book, while the second and third digits offer a more specific and detailed description of the book, providing a more refined classification within the broader subject category. The fiction books are not assigned a Dewey number

### 4.1.3 Capacity Planning

Table 6 shows the shift table for a regular week at the main library. The different areas of the library are categorized as A, B, and C. A describes the information counter on the first floor. In the early shift, one person from A is responsible for the sorting area, handling the incoming books from other branches. B is responsible for the children and youth counter on the first floor and C is responsible for the counter on the second floor. On Fridays, the library closes at 16:00, so there is no late shift. On weekdays, the early shift lasts from 08:30 to 12:00, the mid shift from 12:00 to 15:00, and the late shift starts at 15:00 and ends at 18:15. On Fridays, the mid shift ends at 16:00. During the weekends there is only one shift per day, Saturdays from 10:45 to 16:15, and Sundays from 11:45 to 16:15. The “flex” person on the Saturday shift rotates between A, B, and C, allowing everyone to have a lunch break throughout the day. The employees manning the information counters only work one shift per day. Furthermore, several other employees are working at the library related to administrative tasks, thus not usually servicing patrons, or performing material handling tasks.

| Day / Shift | Early | Mid | Late | Weekend |
|---|---|---|---|---|
| Monday to Thursday | **A:** 3<br>**B:** 1<br>**C:** 2 | **A:** 2<br>**B:** 1<br>**C:** 2 | **A:** 1<br>**B:** 1<br>**C:** 2 | |
| Friday | **A:** 3<br>**B:** 1<br>**C:** 2 | **A:** 2<br>**B:** 2<br>**C:** 2 | | |
| Saturday | | | | **A:** 2<br>**B:** 2<br>**C:** 2<br>**Flex:** 1 |
| Sunday | | | | **A:** 2<br>**B:** 1<br>**C:** 2 |

Table 6: Weekly workforce allocation in the main library.

### 4.1.4 Sorting Machine

The main branch is currently the only branch equipped with a sorting machine. The sorting machine consists of three self-return stations equipped with RFID-readers, a conveyor belt, and multiple bins placed alongside the conveyor belt. Once books have been passed through the self-return machines, the conveyor belt transports the media into the sorting machine. The RFID-tag is scanned by the machines, and a query is sent to the LMS, Bibliofil, which determines which bin the media will be placed in. Once a bin is full, it has to be emptied, and staff moves the material out of the sorting area. As explained in Section 3.1.2 and displayed in Figure 23, the sorting machine produces sorting logs based on the queries and responses generated by each of the three self-return machines connected to the sorting machine. This is the only data available that describes how returns are sorted in the sorting machine, and which bin each returned book is placed into. Furthermore, the files that originate from the self-return machines available to patrons are only saved for four weeks, while the files from the third machine are saved indefinitely.

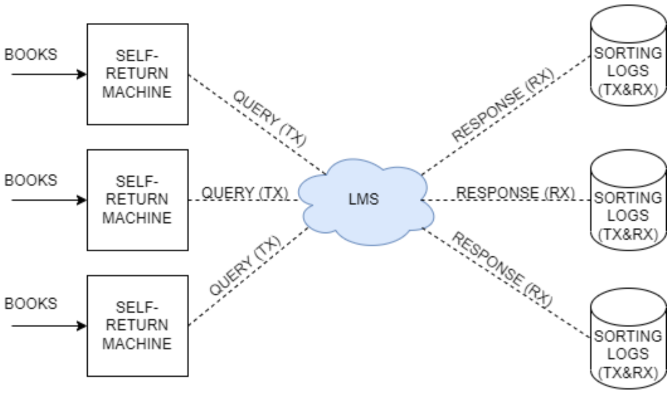


Figure 23: Generation of sorting logs.

The self-return stations labeled B in Figure 24 are available to patrons, while the one labeled C is devoted to staff. The arrows depict the flow through the machine, and the bins, labeled E, are individually enumerated. The two circles in the top right, labeled A, signal when a bin is full by flashing blue. If the left light flashes, it is a bin on the left side of the machine that is full.

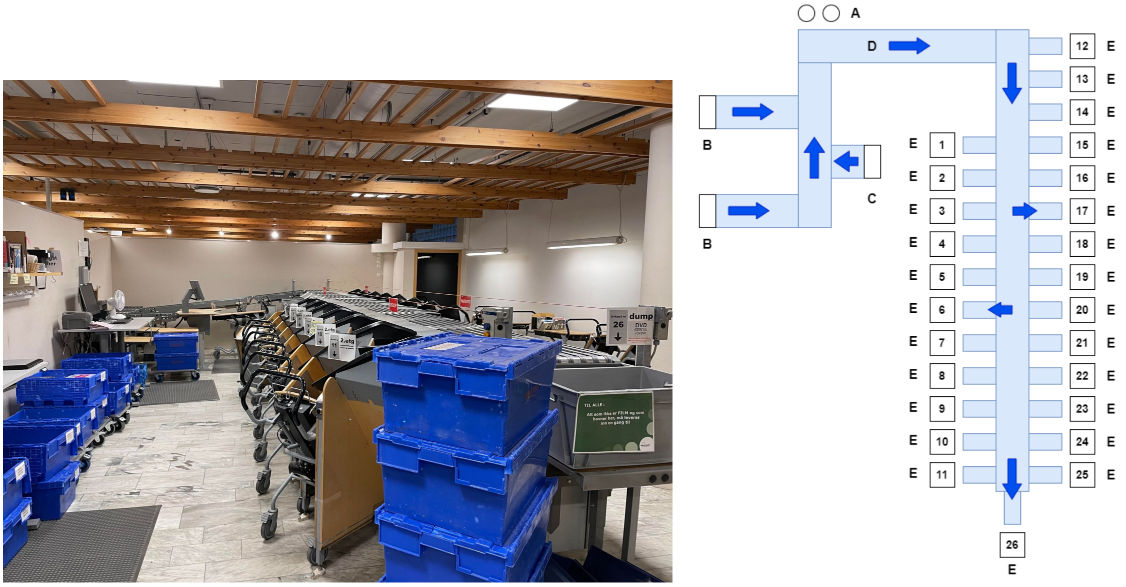
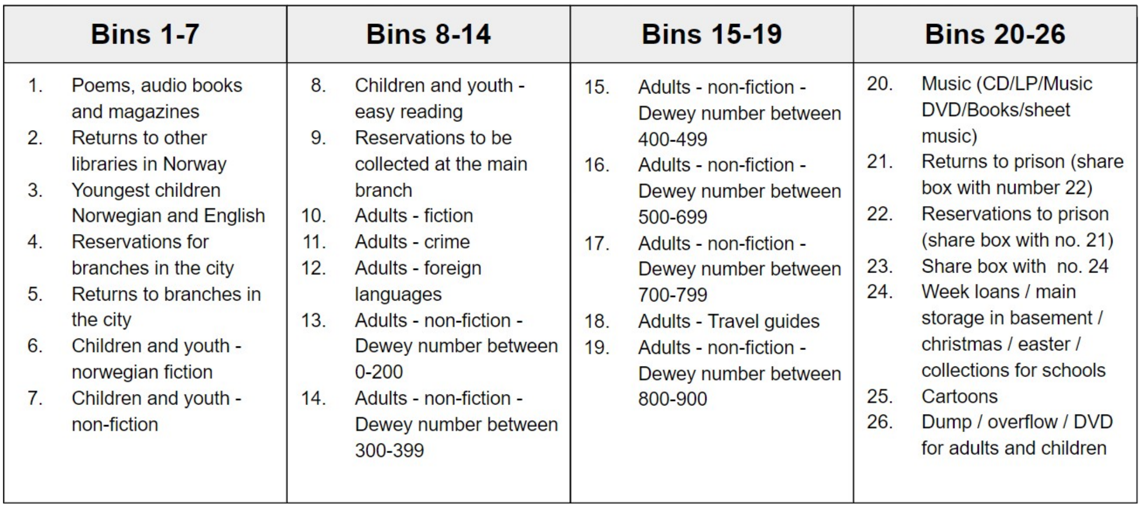


Figure 24: Sorting machine at main library.

Figure 25 displays a description of the current bin allocation at the main library. The bins with percentages refer to bins that are divided over several floors. Books from bins 1, 2, 3, 6, 7, 8, 9, 24, and 25 belong on the first floor of the library. Books from bins 1 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, and 26 belong on the second floor. Additionally, a considerable amount of the books in bin 24 are going to the basement storage, and books in bins 4, 5, 21, and 22 are going to other branches with the truck.



| Bins 1-7 | Bins 8-14 | Bins 15-19 | Bins 20-26 |
|---|---|---|---|
| 1. Poems, audio books and magazines<br>2. Returns to other libraries in Norway<br>3. Youngest children Norwegian and English<br>4. Reservations for branches in the city<br>5. Returns to branches in the city<br>6. Children and youth - norwegian fiction<br>7. Children and youth - non-fiction | 8. Children and youth - easy reading<br>9. Reservations to be collected at the main branch<br>10. Adults - fiction<br>11. Adults - crime<br>12. Adults - foreign languages<br>13. Adults - non-fiction - Dewey number between 0-200<br>14. Adults - non-fiction - Dewey number between 300-399 | 15. Adults - non-fiction - Dewey number between 400-499<br>16. Adults - non-fiction - Dewey number between 500-699<br>17. Adults - non-fiction - Dewey number between 700-799<br>18. Adults - Travel guides<br>19. Adults - non-fiction - Dewey number between 800-900 | 20. Music (CD/LP/Music DVD/Books/sheet music)<br>21. Returns to prison (share box with number 22)<br>22. Reservations to prison (share box with no. 21)<br>23. Share box with no. 24<br>24. Week loans / main storage in basement / christmas / easter / collections for schools<br>25. Cartoons<br>26. Dump / overflow / DVD for adults and children |

Figure 25: Current bin allocation at the main library

The sorting machine in question was procured in 2007 and is positioned on the first floor, directly behind the information desk, which is deemed a crucial area within the library. The staff at TPL have raised multiple concerns regarding the machine, and there is a prevailing notion that it is approaching obsolescence. Apart from its operational and maintenance limitations, the machine also exhibits drawbacks concerning the health and safety of the employees. The drawbacks include:

- Increasingly prone to errors leading to downtime.
- Maintenance is costly.
- Hard to acquire spare parts.
- Software will not be updated.
- It is not possible to expand the machine.
- Limited space around the machine.
- Takes up valuable space in an attractive part of the library and contributes to noise pollution.
- An excess of manual work leads to lifting heavy boxes.

As stated in the introduction, the main goal of the SmartLIB project is to reduce the amount of physical labor performed by staff, enabling staff to devote more time to service patrons. During the specialization project, the sorting machine was highlighted as an integral part of handling incoming material at TPL, as it autonomously sorts approximately 400 000 incoming copies yearly, thus contributing dramatically to minimizing manual work. From the bullet points above, it is clear that a successful sorting machine is detrimental to ensuring that staff can allocate more of their time to service patrons, rather than sorting and moving books. Currently, the sorting machine is the only robotic component of the material handling system at TPL.

## 4.2 Flow of Returned Materials and Related Material Handling

The typical daily responsibilities of a librarian include providing assistance to library patrons, organizing various events, and managing material handling tasks to ensure that returned books are promptly available for new loans. However, in order to enhance the overall value of library services, it is preferable to allocate as much time as possible to the first two activities. Consequently, this case study aims to explore the incoming material flow at the main library, as it initiates several material handling tasks. By examining how the flow of incoming books and the associated material handling activities impact the librarian's workday, valuable insights can be gained.

Information on the processes involved has been gathered through meetings, interviews, and on-site observations. These data sources have provided a comprehensive understanding of how the various steps unfold in the journey from material return to the re-shelving of books in their original locations within the main library. To explain this sequence of steps, a flowchart has been developed.

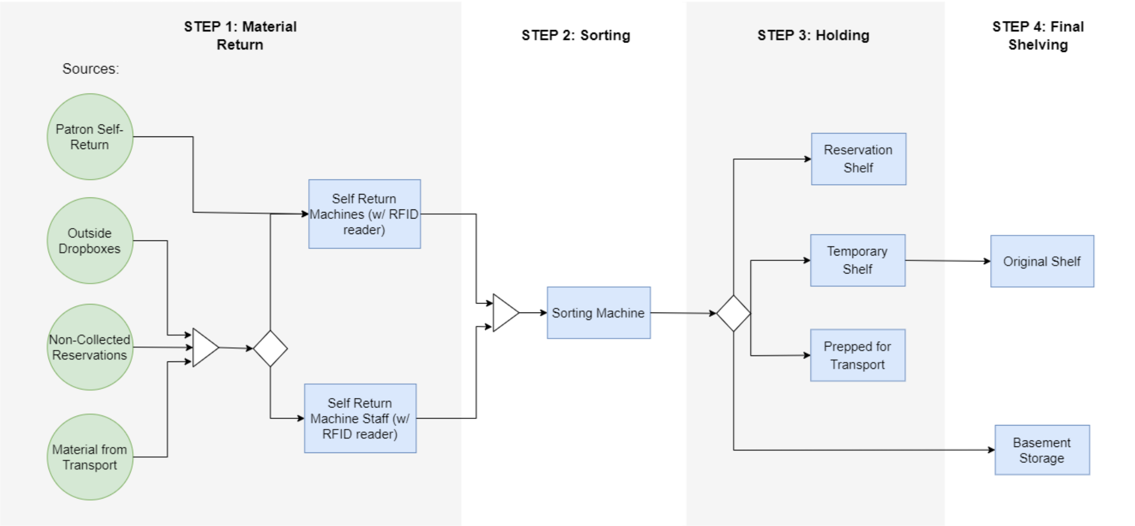


Figure 26: Flowchart explaining the steps from incoming material to material on original shelf.

The time durations presented in the following paragraphs have been derived from data collected through a collaborative effort involving on-site observations conducted by a team of students from the Georgia Institute of Technology, as well as activity logging provided by the library staff. It is important to note that all time estimations provided refer to the duration required for a single employee to complete the entire task. In the subsequent paragraphs, we will elaborate on the results obtained from the data collection process and provide detailed descriptions of the various process steps involved.

### 4.2.1 Step 1: Material Return

As previously mentioned, the main library is equipped with two self-return machines that are accessible to patrons. These machines are directly linked to the sorting machine, facilitating the automatic transportation of materials deposited into them. Consequently, material handling is not required by library staff to handle these returned items. However, patrons also have the option to return books outside of the library's regular operating hours through the utilization of external drop-boxes. These drop-boxes are subsequently collected by staff once the library reopens, and the books are then processed using the self-return machines. Typically, the drop-boxes are collected once prior to closing as well. In total the task necessitates 16 minutes each day, on average, depending on the number of books.

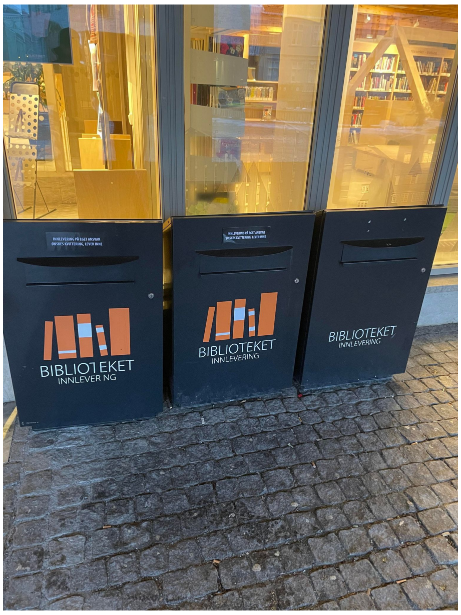


Figure 27: Delivery drop-boxes outside the main entrance

Moreover, in compliance with national regulations, TPL patrons enjoy the convenience of requesting books from any branch within the library network, as well as from other libraries across Norway, and have the possibility to collect the book at their preferred branch. Presently, each book is associated with a specific "owner" branch to which it is returned upon completion of a loan. As a result, numerous books are being transported to, for instance, the main library from the other branches. The books are transported in boxes every single weekday and are collected by staff in the parking lot located in the basement, meaning that no books are to be collected in the basement on Sundays and Mondays. Upon collection, library staff must navigate carts with the boxes through two doorsteps to reach the elevators, which can be a challenging task.

Once on the first floor, the trolleys are brought to the sorting machine where the books are processed, a task that typically consumes a significant amount of time during the early shift. On average, the process of retrieving books from the basement and inputting them into the sorting machine takes approximately 71.1 minutes. However, the duration of this process varies considerably due to fluctuations in the number of books arriving from transport. For instance, Tuesdays tend to have a higher influx of books retrieved from the basement compared to other days. As books are not transported between branches on Saturdays and Sundays, all books accumulated over the weekend arrive at the library on Mondays and are subsequently retrieved on Tuesdays. It is not uncommon for over 800 books to be retrieved from the basement in a single morning, which explains the time-intensive nature of this process

Lastly, an additional source that contributes to the increased flow of books in the sorting machine is the presence of non-collected reservations. When a patron makes a reservation, it initiates a process whereby a librarian must locate the requested book, register it as reserved, and subsequently place it on the reservation shelf or prepare it for transportation to another branch if the patron has chosen an alternate pick-up location. Typically, reservations have a pick-up deadline of six days from the availability notification sent to the patron, except for certain loans that have a 13-day deadline. The extended deadline is applicable to reservations where the availability notification is sent as a physical letter. In some cases, patrons may request a deadline extension by notifying TPL in advance.

Unfortunately, several reserved books are never collected within the specified deadline. At the main library, this results in staff having to retrieve the uncollected books from the reservation shelf and return them using the self-return machines. Removing books from the reservations shelf can be physically demanding for staff. The books are positioned at different heights on the shelf, which often necessitates awkward lifting and can strain the back. Additionally, when books are removed from the shelf, the shelf itself requires sorting. The sorting is conducted chronologically, with the oldest reservation located in the top left position and the newest reservation in the bottom right position. Based on observations, this sorting process typically takes an average of 9.2 minutes to complete.

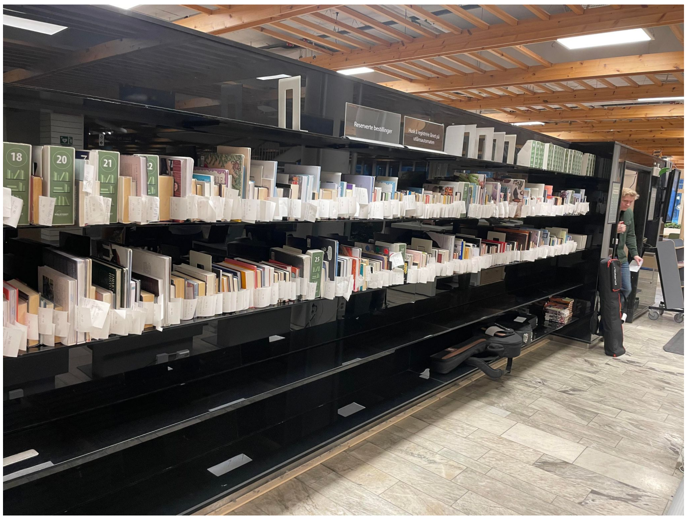


Figure 28: The reservation shelf at the main library.

To summarize, the manual material handling tasks related to managing the material return require 96.3 minutes in total. However, as explained with the boxes retrieved from the basement, these tasks are not necessarily performed each day. Therefore, the average amount of time spent on these activities equates to 80.75 minutes each day.

### 4.2.2 Step 2: Sorting

The sorting process at the main library is predominantly automated, facilitated by the employment of a book sorting machine. To decrease the physical strain on staff members, the bins that are to be moved to other areas in the library are equipped with trolleys. The last bin, bin 26, is assigned to manage overflow. Books are directed here if the sorting machine cannot determine their appropriate destination or if their supposed bin is full. In such cases, staff members are required to retrieve these books and reprocess them through the sorting machine, or manually place them in a bin. However, as this occurrence is infrequent, the time expended on this task is negligible

Technical malfunctions of the sorting machine are perceived as a significant concern and, in the worst-case scenario, may necessitate manual sorting. There have been instances where the sorting machine experienced a full-day downtime, requiring all sorting activities to be carried out manually throughout the day. In general, however, the sorting step at the main library consumes a minimal amount of time in the librarians' workday.

### 4.2.3 Step 3: Holding

During the relocation of materials from the sorting machine, library staff undertake various actions such as placing items on a reservation shelf, arranging them on a temporary shelf, transferring them to the basement storage, or preparing them for transportation to another branch.

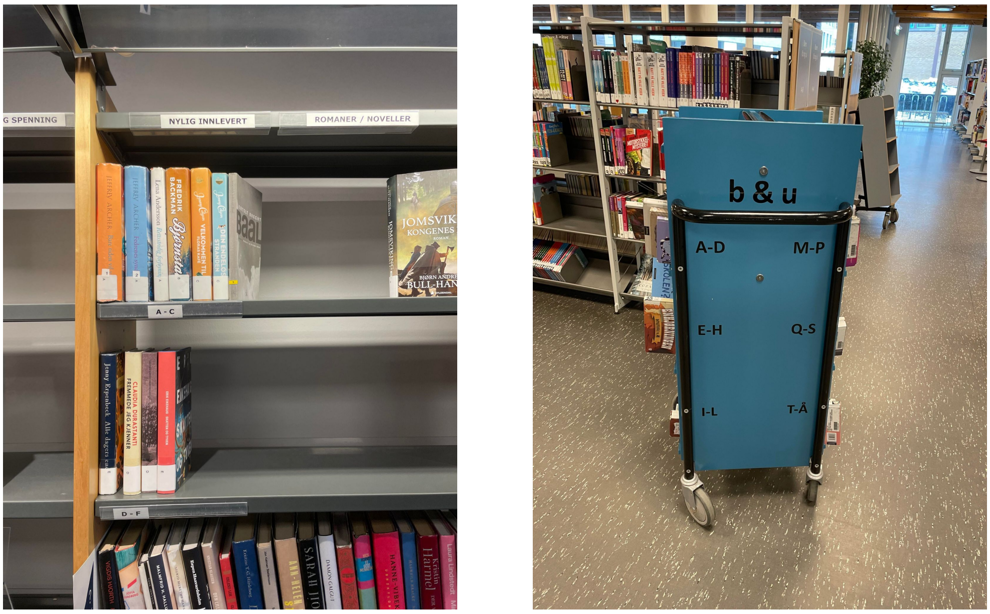
Books that have been reserved for other branches, or need to be returned to their designated branch are placed in bins 4 and 5. These bins are consistently emptied by librarians throughout the day, minimizing the likelihood of reaching maximum capacity. When emptying these bins, librarians are required to scan and manually sort the books into boxes positioned adjacent to the sorting machine, facilitating their subsequent transportation. On average, this process entails a time expenditure of 33.1 minutes per day, varying based on the quantity of books involved. Additionally, books reserved for the main branch that enter the sorting machine are directed to bin 9. This bin is also regularly emptied throughout the day, necessitating librarians in the sorting area to register the reservation as a loan, notify the patron regarding the availability of the book for pick-up, and place the book on the reservation shelf. The average duration for this task amounts to 41.3 minutes each day.

Moreover, bins 23 and 24 are devoted to books that are to be placed in the storage area in the basement. When moving books down to the basement, the books are moved to a second, temporary, trolley, before they are placed in the storage area. On average, the time spent on moving books to the basement and correctly placing them in the storage area is 25 minutes each day.

The final alternative during the holding step involves placing books on temporary shelves. Temporary shelves serve as an intermediate stage in the material handling process to prevent an excessive amount of time being devoted to placing books in their respective areas in the library, thus ensuring that staff does not remain unavailable to assist patrons for longer periods of time. Consequently, temporary shelves are strategically positioned on each floor, with books being allocated to the shelf nearest to their original location.

Emptying the trolleys onto the temporary shelves designated for adult books is a swift task. An elevator in the sorting area specifically facilitates the transportation of trolleys to the second and third floors. The temporary shelves are organized in alphabetical order based on genres or according to the Dewey decimal classification system. However, the books themselves do not need to be arranged in the correct order on the shelves. On average, a combined time of 19.4 minutes is dedicated to transferring full trolleys of adult books to temporary shelves and sorting them into temporary shelves, each day.

However, the temporary shelving is more cumbersome in the children and youth area on the first floor. Due to limitations in the sorting machine's capacity, children and youth books are not adequately sorted during the sorting process. Consequently, these books must be sorted upon arrival at the temporary shelves, as a single trolley typically contains books destined for different temporary shelves. On average, 60 minutes are expended per day on emptying trolleys onto the temporary shelves in the children and youth area.



(a) The temporary shelf in the adult section on the second floor

(b) The temporary shelf in the children's section

Figure 29: The temporary shelves at TPL

To summarize, 178.8 minutes are spent on manual material handling tasks related to holding each day.

### 4.2.4 Step 4: Final Shelving

The final phase of managing the incoming book flow involves the relocation of books from the temporary shelf to their respective original shelves, a process facilitated through the use of trolleys, different from the ones around the sorting machine, but similar to the one illustrated in Figure 30. The library does not enforce a specific rule regarding the duration of a book's stay on the temporary shelf. However, it is customary for each librarian to handle the equivalent of a full trolley of books on a daily basis, typically when they are not engaged in other duties. Multiple librarians are assigned to this task, each responsible for replenishing books in their designated areas. On the second and third floor, it is estimated that it requires approximately 11 minutes to empty one trolley full of books from temporary shelves to original shelves. In these areas of the library, librarians dedicate an average of one hour to this undertaking each day, combined.

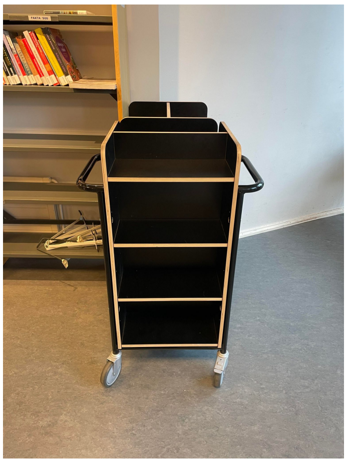


Figure 30: Transportation trolley used for moving books from the temporary shelves to the original shelf.

However, the frequency of emptying the temporary shelves in the children and youth area surpasses that of the adult area on the second and third floors. This discrepancy can be attributed to adults utilizing the temporary shelf for recommendations and showing interest in recently-read materials. Consequently, it is desirable to let books sit on temporary shelves for a longer period of time in these areas. On average, approximately eight temporary shelves in the children and youth area are emptied daily. Through observations, it was determined that emptying a single shelf in this area takes approximately 17 minutes. The children and youth area often experiences high foot traffic due to visits from kindergartens and primary schools, which can hinder a smooth return of books to their original shelves. Additionally, children are more inclined to pick books from shelves and place them in random locations. To mitigate this issue, TPL has implemented designated drop-off shelves where children can deposit books, which are then emptied and returned to their proper positions on the original shelves during each shift, often simultaneously with emptying the temporary shelves. Consequently, librarians in the children and youth area spend a cumulative total exceeding 143 minutes per day in the process of restocking the original shelves. However, it is essential to acknowledge that during restocking, librarians may also engage in concurrent activities, such as assisting patrons or organizing the shelves, thereby contributing to the overall time expended on this task.

In total, approximately 203 minutes are spent on moving books from temporary to original shelves on a daily basis.

### 4.2.5 Time Spent on Material Handling

After describing the processes related to managing the incoming book flow, and presenting an overview of time spent on the different steps, it can be established that librarians at TPL spend a considerable amount of time performing manual material handling tasks each day. From the data describing time spent on these manual material handling tasks, both logged by the librarians in the period 23.01.23 - 04.02.23, and derived through observations at the library, it has been calculated that approximately 5550 minutes were spent on manual material handling activities in the period. This averages to the library staff spending approximately 462.25 minutes on manual work related to managing the incoming book flow daily. This number, however, is an estimate that will vary daily based on factors such as the number of incoming books and reservations, but was observed to be mainly dependent on weekdays and weekends. The amount of time spent on manual work fluctuated in the period as follows:

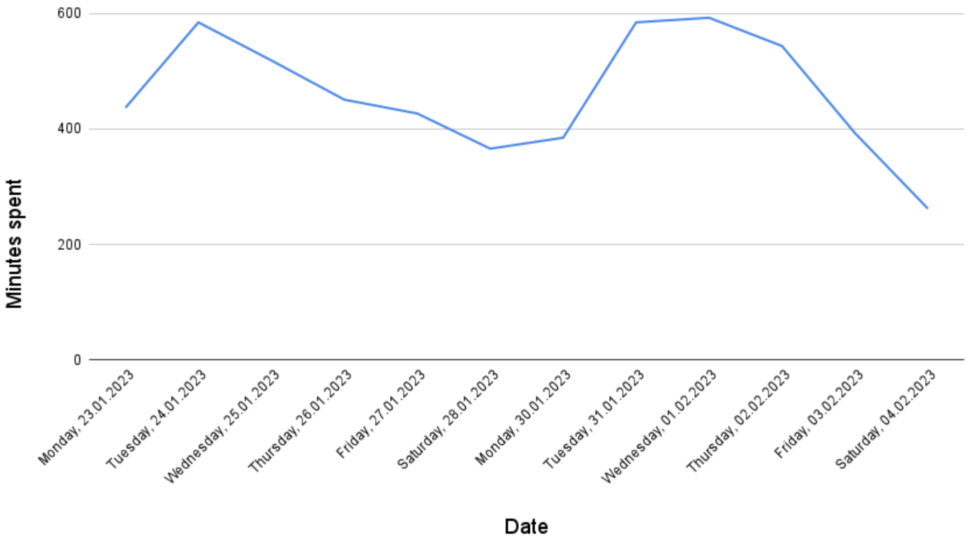


Figure 31: Distribution of minutes used on manual material handling by day.

### Material Handling in Relation to Day and Shift

It is clear that more time is spent during weekdays, especially midweek. However, fewer employees are working during weekends. The following graph depicts how the amount of manual material handling work corresponds to the manhours available for each day:

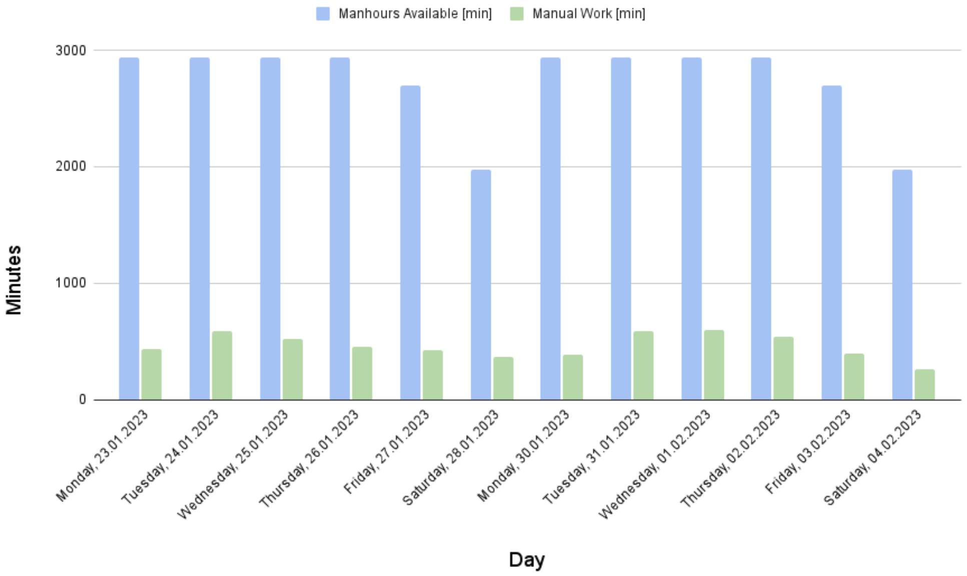


Figure 32: Time spent on manual material handling tasks compared to available manhours.

The manhours available are derived from the staffing list displayed earlier in this section. For Saturdays, one employee, displayed as “flex” in the staffing list, has been removed from the available manhours, as it is a continuously rotating position, where the employees are having a break. Therefore, these have been excluded from available manhours. From the figure, it becomes apparent that approximately 17 percent of the librarians’ workday was spent performing manual material handling tasks, on average in this period. The deviation from the average in this period is low. Moreover, by analyzing the data it was possible to derive the amount of time spent on manual material handling tasks on average each shift, and how this corresponds to the total time available on each shift:

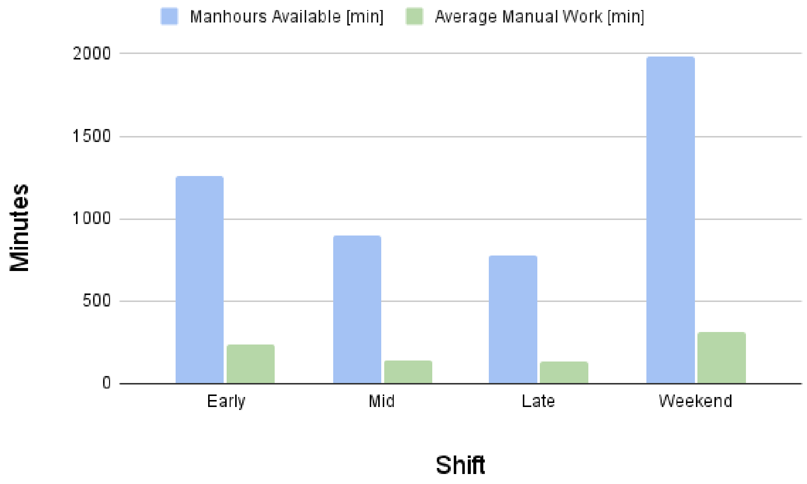


Figure 33: Distribution of minutes used on manual material handling by shift.

In the early shift, around 19 percent of available manhours were spent on conducting manual material handling tasks, in comparison to 16 percent and 17.5 percent on the mid and late shifts, respectively. During weekends, 17 percent of the available time was spent on manual material handling tasks.

### Time spent in Relation to Task and Step

The material handling tasks that was included in this case study, and have constituted the total amount of time, are as follows:

| Number | Manual material handling task |
|---|---|
| **1** | All manual work related to non-collected reservations. |
| **2** | All manual work related to returns in the outdoor boxes. |
| **3** | All manual work related to getting full boxes from the garage and returning the books to the sorting machine. |
| **4** | Reservations to be sent to other branches in the city and other Norwegian libraries. |
| **5** | Reservations to be collected at the Main branch incl. the time it takes to put in on the reservation shelf. |
| **6** | Moving full trolleys to the temporary shelf at the adult department at the main branch. |
| **7** | Moving full trolleys to the temporary shelf at the youth and child department at the main branch. |
| **8** | Moving books to the original shelf. |
| **9** | Moving books to the basement. |

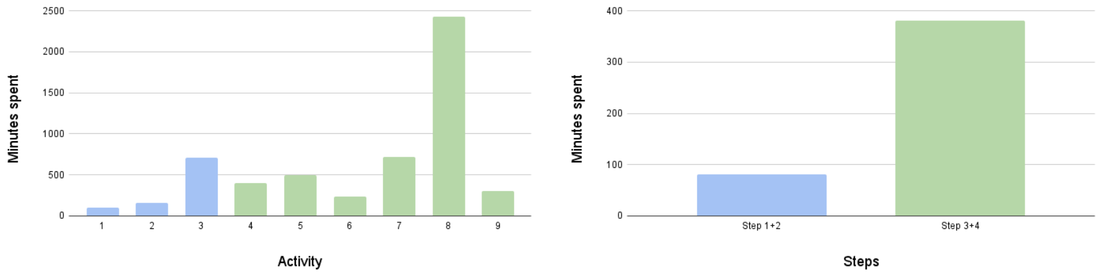
Table 7: Description of the observed manual material handling tasks.

It has been observed that tasks related to the material return and sorting usually are performed by the two librarians designated to the first-floor information desk and the sorting area, alongside the employee responsible for managing the book arriving from transport. These material handling tasks are usually performed in quick succession as close to opening hours as possible, which was found to be beneficial for several reasons:

- Less crowded at the library during opening hours, which implies less demand for servicing patrons.
- The library is able to restock its shelves in the morning, particularly beneficial in the children and youth area.
- Reservations are handled as quickly as possible.

The tasks related to the holding and final shelving, however, are normally more evenly distributed. Especially the final shelving was expressed to usually be carried out when the librarians find time for it. The librarians responsible for the second and third floors, in addition to the children and youth area, are normally occupied by other tasks, such as vacating the information desks and servicing patrons. Therefore, there is no hurry to place books on their final shelves, as other value-adding tasks are of higher priority. Therefore, minimizing the cycle time of a book in the system by decreasing this time, is not necessarily indicative of optimal performance.

While it was revealed that considerably more time is spent on tasks related to holding and final shelving, compared to managing the material return, it should be mentioned that three employees perform all the material handling related to material return, while nine employees perform the tasks on the last steps. As a result, the amount of manual work per employee is relatively equal.



(a) Total minutes spent on each activity

(b) Minutes spent on average on the steps before and after the sorting machine

Figure 34: Minutes spent on the different activities and steps.

### Relationship Between Incoming Book Flow and Material Handling

Further, by analyzing the volume of the incoming book flow and the time spent on material handling each day, it is possible to examine the relationship between these two variables. To examine the relationship between the book flow and time spent on manual material handling tasks the Pearson correlation coefficient was used:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{2}$$

The correlation is based on the time spent on material handling each day versus the volume of the incoming book flow at the main branch per day. This resulted in a correlation of 0.762, which is classified as a high positive correlation (Asuero et al. 2006). Awareness of the incoming book flow might therefore be thought to be closely related to capacity planning. Currently, capacity planning is static and loosely based on previous experiences, as depicted in Section 4.1.3. Figure 35 provides a visual representation of the relationship between delivered books at the main library plus books prepared for reservation, and minutes spent on manual material handling tasks.

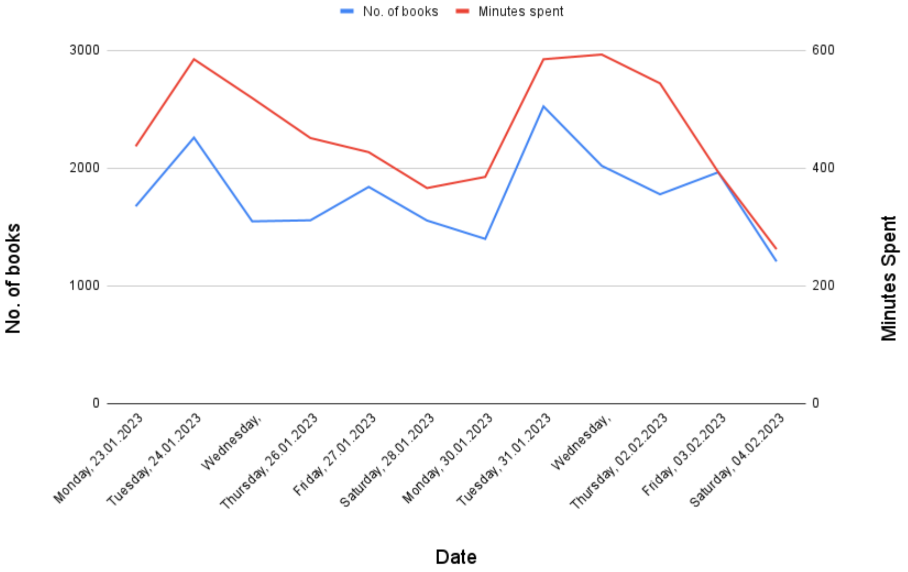


Figure 35: Books delivered and prepared for reservation plotted with minutes spent on manual work.

### 4.2.6 The Process from an Ergonomic Perspective

From an ergonomic standpoint, there are multiple flaws in the current operations at TPL, which were discovered through observations and discussions with the librarians. Firstly, boxes with books are generally heavy and awkward to lift. This is the case for the outside drop-boxes and the boxes retrieved from the basement. Secondly, a large portion of shelves is at awkward heights for the librarians. This is especially the case in the children and youth area, as the shelves are adapted for the intended patrons, and not the librarians. Therefore, it can be straining for librarians to move and sort books in this area, but also in the lowest part of all the other shelves. This is unfortunate, as it is an operation that is performed numerous times throughout the day. Moreover, multiple employees have expressed dissatisfaction with the reservation shelf. This shelf is cumbersome to manage with the current system; as the old, non-collected, reservations are removed from the top left, every single book has to be moved toward the top left. This is especially challenging when books have to be moved from a low shelf. Furthermore, the librarians stationed in and around the sorting machine experience the loud noise from this machine constantly and has to be alert for full shelves at all times, which can be stressful. Also, as mentioned earlier, it makes employing the information desk incoherent where the employees have to let go of their working tasks whenever a bin is full.

One of the top priorities at TPL is the employees' well-being and health, and according to the head of IT, the library is currently experiencing too much sick leave. However, large changes and alterations to the current processes are unrealistic, as it is costly and resource-consuming. Changes, such as the introduction of mobile robots would likely be implemented gradually and with a long time horizon. It is therefore important to aim to minimize the amount of physical work performed by librarians with the current system.

## 4.3 Challenges Present in the Current Sorting System

After examining the material handling necessary to manage the incoming book flow to the main library of TPL, the sorting machine can be stated to be an integral part of the material handling system. Firstly, the sorting machine automates the sorting processes at the library. Secondly, it can be argued that sorting affects other aspects of the material handling system, especially related to holding. These tasks include:

- Emptying bins around the sorting machine and temporary shelving.
- Preparing books for reservations.
- Preparing books for transport to other branches/libraries.

Consequently, finding an optimal setup in the sorting machine will contribute to decreasing the amount of time staff will have to devote to material handling, as a considerable amount of time is spent on holding tasks each day. An interesting aspect is examining and reviewing the allocation of bins in the sorting machine. Currently, the allocation of bins is static and determined by Bibliofil, as presented in Figure 25. This allocation will from here on be referred to as the current allocation, or the AS-IS.

By examining historical data, it is possible to derive the probabilities of which bin a book will be placed in. The bins are believed to have a capacity of +/-22 books, and approximately 1400 books are passed through the sorting machine on average each day. The total distribution of books in the different bins in the sample period is depicted below:

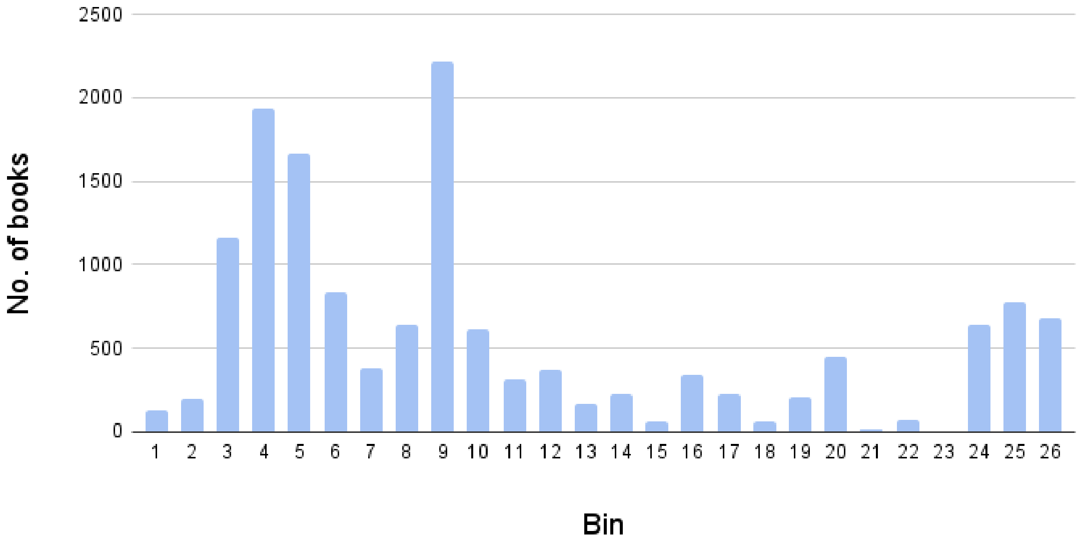


Figure 36: Books going to the different bins in the period: 23.01.23 - 04.02.23

The subsequent sections will present issues identified in the current sorting process.

### 4.3.1 Uneven Fill Rates

As illustrated in Figure 36, the sorting machine is currently characterized by uneven fill rates in the different bins. In theory, this should magnify the occurrence of bins having to be refilled, as demonstrated in Figure 37, which results in additional manual work being done by librarians. Furthermore, library staff working in the sorting area have expressed dissatisfaction about the uneven fill rates; bins are filled sporadically throughout the day at an inconsistent tempo, meaning that staff will have to work accordingly. It can therefore be considered as a source of additional stress for the library staff.

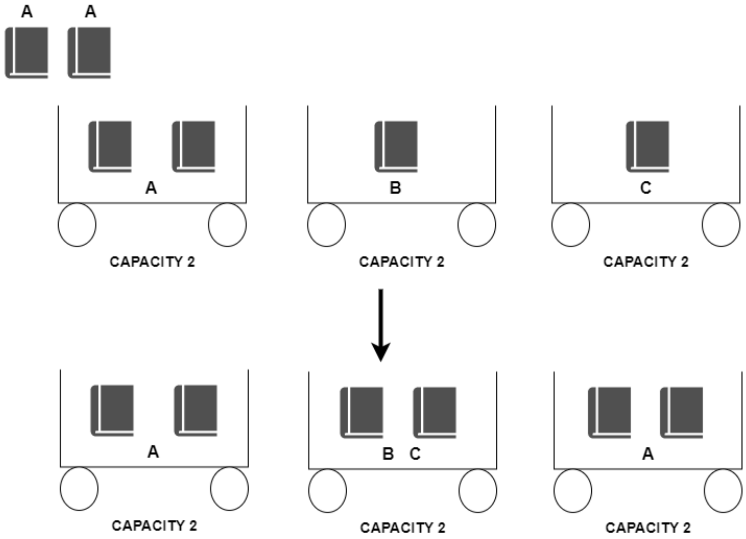


Figure 37: Demonstration on levelling fill rates in the sorting machine

Figure 38 illustrates how often the different bins were filled each day in the examined period. The blue dots in the diagram represents the number of times a bin was filled each day. The red dots symbolize the median number of times a bin was filled daily, while the green dots symbolize the average each day for a bin. The diagram highlights the uneven fill rates, with bin 9 being filled up to 13 times a single day, while several bins were not even filled once on multiple occasions. However, what is not displayed by the figure, is when the bins had to be refilled. As explained previously, a large number of books enter the sorting machine in the beginning of each day, when books from transport are retrieved from the basement, non-collected reservations are handled, and the outside drop-boxes are emptied. Naturally, bins are filled at a high tempo during these processes, resulting in intense hours in the sorting area during the first hours after opening.

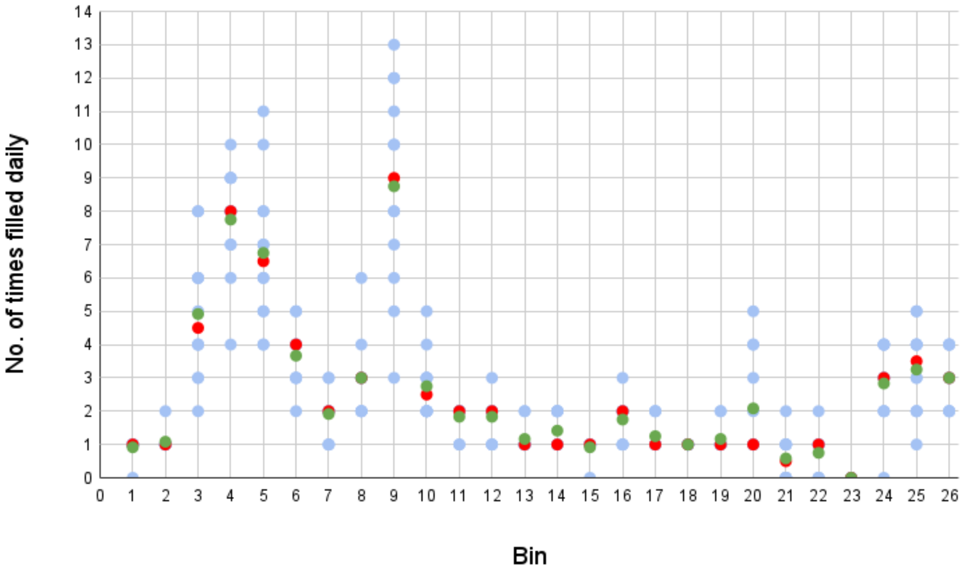


Figure 38: Number of times bins filled per day in the period: 23.01.23 - 04.02.23

Furthermore, it is clear that the fill rates of the bins vary from day to day. Table 8 provides additional information and statistics about the fill rates of the bins. From the table, it becomes apparent that there were a total of 15 bins that have a median number of times filled each day equalling 0 or 1. Only two of these 15 bins were filled more than twice in a single day during the examined two-week period. As a result, the number of times these bins are filled in a day is relatively stable. The fill rates for the bins that receive more books, however, are characterized by a greater deviation and variance, adding a level of unpredictability to the demand for these bins. This indicates that a dynamic allocation of bins might be beneficial for the system. A general rule of storage systems states that a more dedicated and static approach is desirable if the input flow to the system is predictable, while a more dynamic approach is optimal if the input flow of items is unpredictable (Berg 1999, Roodbergen and Meller 2004, Bakkali et al. 2013). Furthermore, the bins that are filled less frequently can also be seen as a loss of space around the sorting machine.

| **Bin** | **Min/Max** | **Median** | **Average** | **Standard Deviation** | **Variance** |
|---|---|---|---|---|---|
| 1 | 0/1 | 1 | 0,9167 | 0,2887 | 0,0833 |
| 2 | 1/2 | 1 | 1,0833 | 0,2887 | 0,0833 |
| 3 | 2/8 | 4,5 | 4,9167 | 1,9287 | 3,7197 |
| 4 | 4/10 | 8 | 7,7500 | 1,6026 | 2,5682 |
| 5 | 4/11 | 6,5 | 6,7500 | 2,2208 | 4,9318 |
| 6 | 2/5 | 4 | 3,6667 | 0,8876 | 0,7879 |
| 7 | 1/3 | 2 | 1,9167 | 0,6686 | 0,4470 |
| 8 | 2/6 | 3 | 3,0000 | 1,1282 | 1,2727 |
| 9 | 3/13 | 9 | 8,7500 | 3,1079 | 9,6591 |
| 10 | 2/5 | 2,5 | 2,7500 | 0,9653 | 0,9318 |
| 11 | 1/2 | 2 | 1,8333 | 0,3892 | 0,1515 |
| 12 | 1/3 | 2 | 1,8333 | 0,5774 | 0,3333 |
| 13 | 1/2 | 1 | 1,1667 | 0,3892 | 0,1515 |
| 14 | 1/2 | 1 | 1,4167 | 0,5149 | 0,2652 |
| 15 | 0/1 | 1 | 0,9167 | 0,2887 | 0,0833 |
| 16 | 1/3 | 2 | 1,7500 | 0,6216 | 0,3864 |
| 17 | 1/2 | 1 | 1,2500 | 0,4523 | 0,2045 |
| 18 | 1/1 | 1 | 1,0000 | 0,0000 | 0,0000 |
| 19 | 1/2 | 1 | 1,1667 | 0,3892 | 0,1515 |
| 20 | 1/5 | 1 | 2,0833 | 1,5050 | 2,2652 |
| 21 | 0/2 | 0,5 | 0,5833 | 0,6686 | 0,4470 |
| 22 | 0/2 | 1 | 0,7500 | 0,6216 | 0,3864 |
| 23 | 0/0 | 0 | 0,0000 | 0,0000 | 0,0000 |
| 24 | 0/4 | 3 | 2,8333 | 1,1934 | 1,4242 |
| 25 | 1/5 | 3,5 | 3,2500 | 1,2881 | 1,6591 |
| 26 | 2/4 | 3 | 3,0000 | 0,8528 | 0,7273 |

Table 8: Min/Max, Median, Average, Variance, and Standard Deviation of the fill rates to each bin with the current allocation.

### 4.3.2 Sorting of Books Going to Other Branches

Books from other branches constitute a large part of the incoming book flow at the main library, as exemplified by the observed time spent managing books retrieved from the basement. However, there is also a considerable amount of outgoing books at the main library, destined for other branches. In fact, in excess of 30 percent of the books that arrived in the sorting machine from 23.01.23 - 04.02.23 were prepped for transport to other branches. This equates to approximately 350 books on a daily basis. Once these books arrive in the sorting machine, staff has to scan each copy, and manually sort the books in bins designated for transportation to the other branches. This is a repetitive material handling task that could be reduced or avoided with a different setup in the sorting machine.

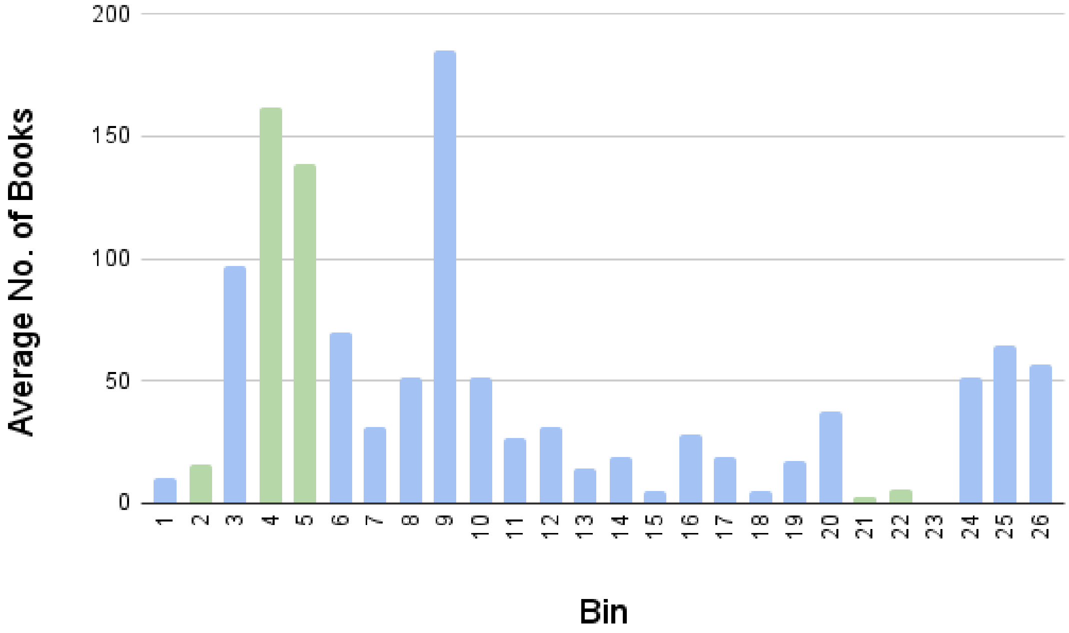


Figure 39: Books prepped for further transport in the sorting area.

Currently, only bins 21 and 22 are devoted to books destined for the prison library, while 4 and 5 are devoted to the other eight branches in the city. This is an odd prioritization, and in the examined time period, an average of 300 books entered bins 4 and 5 each day, whereas just an average of nine books entered bins 21 and 22 each day. Bin 2 is assigned to books being sent to other libraries in Norway. On average, 16 books were sent to this bin every day.

### 4.3.3 Sorting of Children's Books

It was previously described how the sorting of children's books is a problem for TPL, and how it affects the manual work related to placing the books on temporary- and original shelves. There are four bins devoted to children's books in the sorting machine:

- Bin 3: Youngest children, Norwegian and English.
- Bin 6: Children and youth - Norwegian fiction.
- Bin 7: Children and youth - non-fiction.
- Bin 8: Children and youth - easy reading.

Currently, there are six temporary shelves for children's books. A problem with the current bin allocation is that the books that are sorted in the same bin, do not necessarily belong on the same temporary shelf. The books coming from one trolley could belong to up to three different temporary shelves. This makes it cumbersome for librarians to empty the trolleys from the sorting machine in the children and youth area. It is estimated that it requires approximately 4.5 minutes to empty a trolley from the sorting machine into the temporary shelvesFigure 40 illustrates the portion of material handling tasks related to holding that was spent on temporary shelving of children's books:

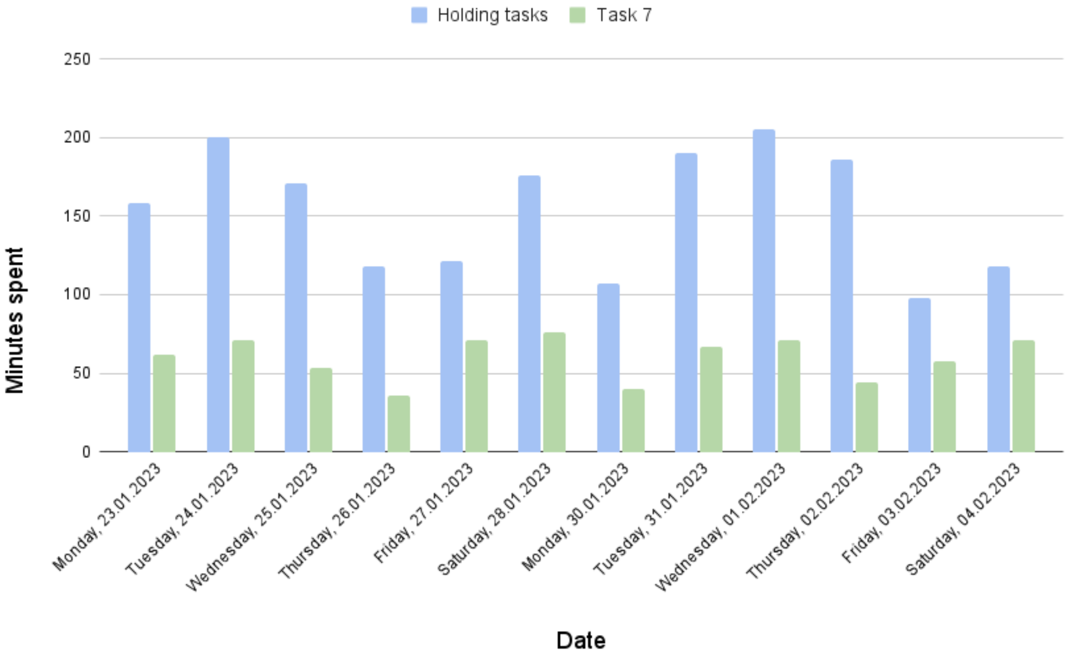


Figure 40: Amount of material handling related to moving children's books to temporary shelves compared with total amount of material handling related to the Holding tasks.

The total amount of time spent on moving children's books to temporary shelves constituted on average 39% of the total time spent on material handling related to the holding tasks, while the incoming book flow only was constituted of 21% children's books.

Moreover, there are currently ten bins in the sorting machine devoted to the sorting of adult's books. The uneven distribution of bins devoted to children's and adult's books seems illogical and unjustified, considering that 3012 books were sorted in the four bins designated for children's books, in comparison to 2609 books sorted in the ten bins designated for adult's books in the period 23.01 to 04.02.

## 4.4 Effects on Material Handling by Altering the Sorting System

After establishing potential areas of improvement in the sorting machine, an obvious continuation is to look for alternative ways of allocating the bins. There are multiple possibilities, that also can be used in combination:

- Combining bins. This can be done to make space for bins that are frequently filled.
- Dividing bins. Used to spread the demand across multiple bins.
- Alter the sorting by creating new sorting criteria.

The following paragraphs will present possible alterations, and how they impact manual work and utility rates. Some constraints have been considered when exploring a new bin allocation, based on discussions with the librarians:

- Do not combine bins with books going to "very" different parts of the library. This would be counter-effective from an efficiency standpoint.
- Keep the overflow bin.

The following paragraphs will present three possible setups of the sorting machine, derived by focusing on leveling the fill rates in the machine, assigning more bins to books being prepped for transport to other branches, or more detailed sorting of children's books.

**1. Leveling the Fill Rates:**
There are multiple layouts that can be derived by combining bins with a low utilization rate and dividing bins with a high utilization rate, thus leveling the fill rates. In the proposal below, the following bins have been combined:

**Bins 2 and 4 → Bin 2**
Bin 2 is devoted to books that shall be returned to other libraries in Norway, while bin 4 is devoted to the reservations for TPL's branches. These bins are similar because the books placed in them will have to be manually sorted and placed in transport boxes once they are retrieved from the bin. Therefore, it could make sense to combine these bins to free up space, even though it does not directly contribute to even the fill rates.

**Bins 13 and 15 → Bin 13**
These bins are filled on average approximately once a day, with the median for the examined period equalling one for both of these bins. Furthermore, the books belonging to these bins are all categorized as "Adults - non-fiction" and have similar Dewey numbers, which means they are placed on the same temporary shelf. It is therefore unwarranted that these bins occupy two spots in the sorting machine.

**Bins 21 and 22 → Bin 21**
Combining bins 21 and 22 - These bins are used to sort books going to the prison, and are filled slightly above once each day, combined. All books going to the prison are placed in the same transport box in the sorting area, it will therefore not require a noteworthy amount of extra work to handle these bins combined.

**Bins 23 and 24 → Bin 24**
Combining bins 23 and 24 - Currently, these bins share one box. Given the relatively low utilization of these two bins combined, it would be smart to prioritize space for other bins, by combining bin 23 and bin 24.

These alterations have made five bins available. Logically, expanding the most frequently filled bins seems like the best choice. These are bins 4, 5, and 9. The setup in the sorting machine is presented in Figure 41, where the bins marked in purple illustrate have been merged, while the green, yellow, and red bins have been divided.

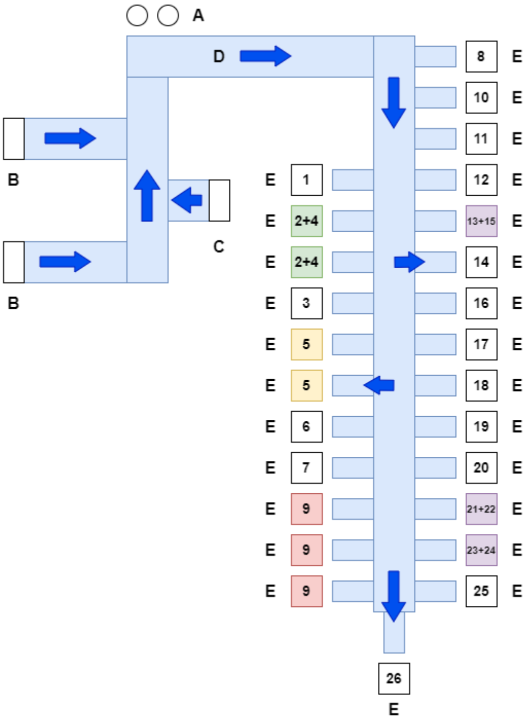


Figure 41: Bin setup in the sorting machine focusing on leveling the fill rates.

Such an alteration of the sorting machine would generate the following statistic in the same period as Table 8:

| **Bin** | **Min/Max** | **Median** | **Average** | **Standard Deviation** | **Variance** |
|---|---|---|---|---|---|
| 1 | 0/1 | 1 | 0,9167 | 0,2887 | 0,0833 |
| 2 | 3/6 | 4 | 4,4167 | 0,7930 | 0,6288 |
| 3 | 2/8 | 4,5 | 4,9167 | 1,9287 | 3,7197 |
| 4 | 3/6 | 4 | 4,4167 | 0,7930 | 0,6288 |
| 5 | 2/6 | 3,5 | 3,5833 | 1,1645 | 1,3561 |
| 6 | 2/5 | 4 | 3,6667 | 0,8876 | 0,7879 |
| 7 | 1/3 | 2 | 1,9167 | 0,6686 | 0,4470 |
| 8 | 2/6 | 3 | 3,0000 | 1,1282 | 1,2727 |
| 9 | 1/5 | 3,5 | 3,2500 | 1,1382 | 1,2955 |
| 10 | 2/5 | 2,5 | 2,7500 | 0,9653 | 0,9318 |
| 11 | 1/2 | 2 | 1,8333 | 0,3892 | 0,1515 |
| 12 | 1/3 | 2 | 1,8333 | 0,5774 | 0,3333 |
| 13 | 1/2 | 1 | 1,3333 | 0,4924 | 0,2424 |
| 14 | 1/2 | 1 | 1,4167 | 0,5149 | 0,2652 |
| 15 | 2/6 | 3,5 | 3,5833 | 1,1645 | 1,3561 |
| 16 | 1/3 | 2 | 1,7500 | 0,6216 | 0,3864 |
| 17 | 1/2 | 1 | 1,2500 | 0,4523 | 0,2045 |
| 18 | 1/1 | 1 | 1,0000 | 0,0000 | 0,0000 |
| 19 | 1/2 | 1 | 1,1667 | 0,3892 | 0,1515 |
| 20 | 1/5 | 1 | 2,0833 | 1,5050 | 2,2652 |
| 21 | 0/2 | 1 | 0,9167 | 0,6686 | 0,4470 |
| 22 | 1/5 | 3,5 | 3,2500 | 1,1382 | 1,2955 |
| 23 | 1/5 | 3,5 | 3,2500 | 1,1382 | 1,2955 |
| 24 | 0/4 | 3 | 2,8333 | 1,1934 | 1,4242 |
| 25 | 1/5 | 3,5 | 3,2500 | 1,2881 | 1,6591 |
| 26 | 2/4 | 3 | 3,0000 | 0,8528 | 0,7273 |

Table 9: Min/Max, Median, Average, Variance, and Standard Deviation of the fill rates to each bin with the new proposed allocation when splitting frequently filled bins.

Figure 42 depicts how the allocation compares to the current one, with regards to the median number of times a bin is filled each day. The previously skewed distribution of books has definitely been more evenly distributed.

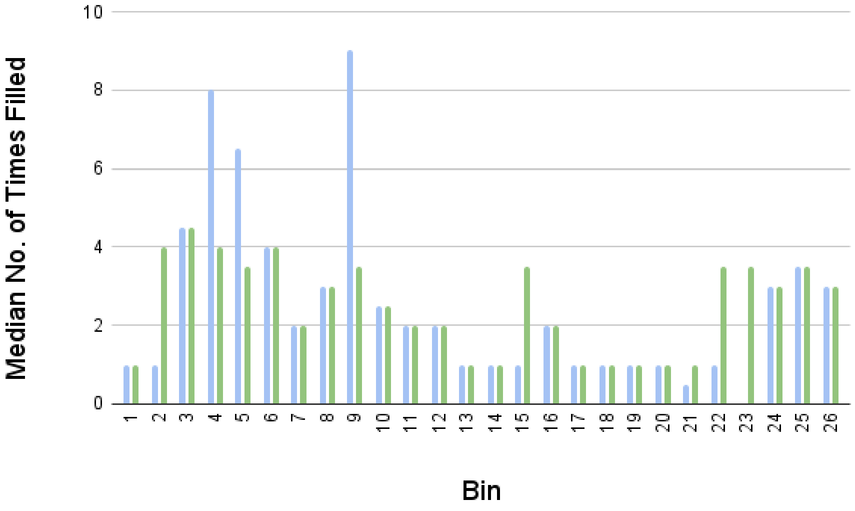


Figure 42: Effect on the median bins filled with the proposed setup.

However, as first believed, the new setup did not affect the number of full bins drastically, and it actually increased from 796 to 799. Nonetheless, there could still be possible savings in time spent on manual work. The expansion of bins provides high-demand bins with higher capacity, as they are assigned to more spaces in the machine. As a result, even though one of the bins is filled, a librarian would not need to empty the bin straight away. Instead, the librarian would have the possibility to wait until another bin is filled, and move both bins to temporary shelves simultaneously, thus saving time. In fact, theoretically, 15 trips could be saved each day.

However, this would not be the case at TPL; all of the bins that are filled with a high frequency, and therefore make sense to divide, are devoted to books that never leave the sorting area. All of these books are either prepared for transport, or placed on the reservation shelf. These bins are emptied continuously, and the books are scanned in the sorting area, therefore no trips to temporary shelves are saved. Nonetheless, more leveled fill rates would propose the librarians the possibility to perform scanning and prepping tasks when desired, and not have to worry about full bins and books being sent to the overflow bin.

### 2. Devote One Individual Bin to Each Branch:

There are multiple possible layouts of bins that would enhance the machine, making it possible to devote one individual bin to each branch. However, to acquire space the necessary space, some bins in the current setup would have to be combined. Currently, bins 2, 4, 5, 21, and 22 are available, as these are already assigned to books to other branches. It is obviously smart to combine the bins with the lowest utilization rates, as long as constraints are satisfied. One of them is presented below, where the following bins have been combined:

**Bins 1 and 18 → Bin 1**

Bin 1 is devoted to poems, audiobooks, and magazines, while bin 18 is used for travel guides. The media entering these two bins are located in close proximity to each other at the library, on the second floor. Therefore, it could make sense to combine these bins to free up space. Furthermore, both bins have a relatively low fill rate, being filled on average once daily.

**Bins 13, 14, and 15 → Bin 13**

These bins are filled on average approximately once a day, with the median for the examined period equalling one for all three of these bins. Furthermore, the books belonging to these bins are all categorized as “Adults - non-fiction” and have similar Dewey numbers, which means they are placed on the same temporary shelf. It will therefore require no additional sorting when placing the books in the temporary shelf.

**Bins 16 and 17 → Bin 17**

These bins are also used to sort “Adults - non-fiction”, and are filled between one and two times each day. These books have similar Dewey numbers and are placed on the same temporary shelf.

**Bins 23 and 24 → Bin 23**

Combining bins 23 and 24 - Currently, these bins share one box. Given the relatively low utilization of these two bins combined, it would be smart to prioritize space for other bins.

These alterations made five bins available, which is the number of bins needed to devote one bin to other libraries in Norway, as well as to each branch in TPL, including the bin already devoted to reservations in the main library. Assigning a separate bin for each branch produces a similar outcome to the leveling of fill rates, as bins 4 and 5, which are frequently filled, experience a change in distribution across eight bins rather than just two with the updated configuration. In Figure 43, the bins that were merged are marked in purple, and bins marked in green illustrate the bins designated for each branch. This alteration would eliminate the need for scanning and sorting the books that are to be further transported in the sorting area, which would on average save 33.1 minutes each day that are spent on these tasks, as quoted previously.

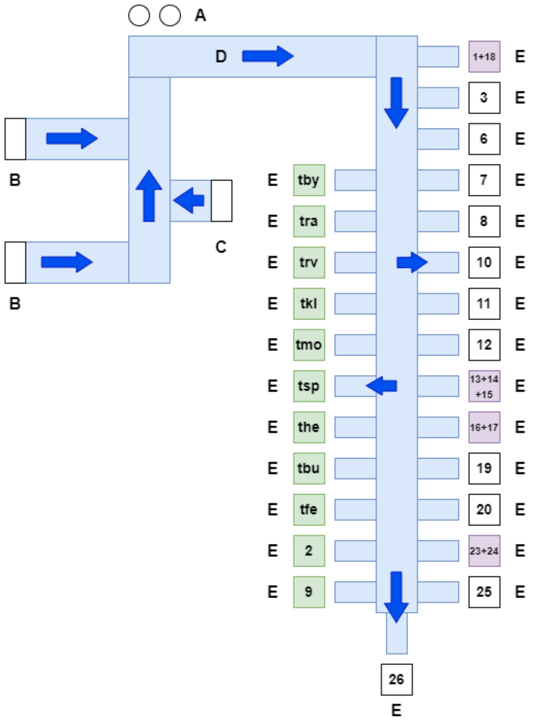
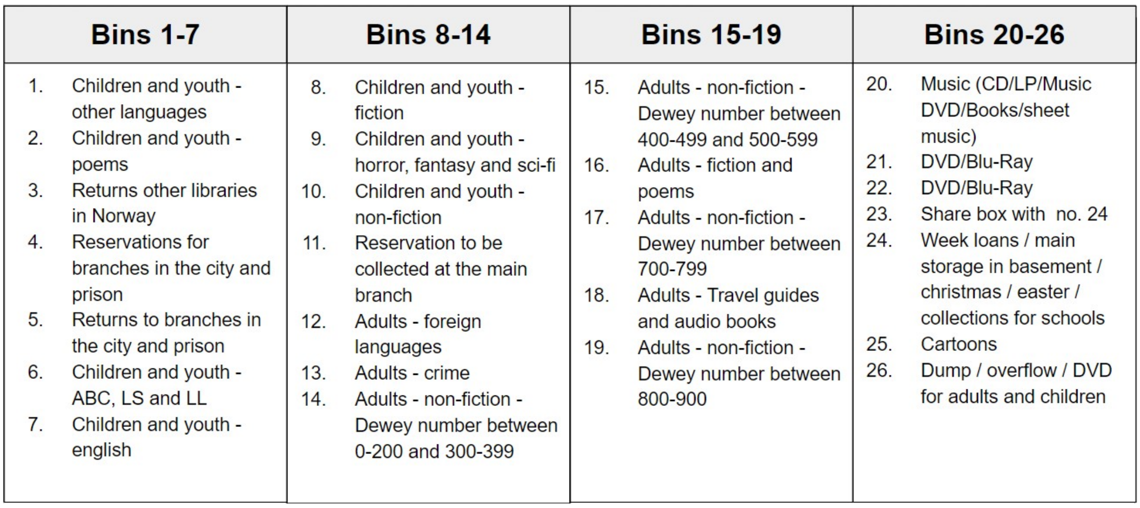


Figure 43: Bin setup in the sorting machine with one bin designated for each branch.

### 3. More Detailed Sorting for Children's Books:

A bin layout that focuses on more detailed sorting for children's books will contribute to streamlining the process of sorting children's books on temporary shelves. Recently, TPL has experimented with such a bin allocation:



| Bins 1-7 | Bins 8-14 | Bins 15-19 | Bins 20-26 |
|---|---|---|---|
| 1. Children and youth - other languages<br>2. Children and youth - poems<br>3. Returns other libraries in Norway<br>4. Reservations for branches in the city and prison<br>5. Returns to branches in the city and prison<br>6. Children and youth - ABC, LS and LL<br>7. Children and youth - english | 8. Children and youth - fiction<br>9. Children and youth - horror, fantasy and sci-fi<br>10. Children and youth - non-fiction<br>11. Reservation to be collected at the main branch<br>12. Adults - foreign languages<br>13. Adults - crime<br>14. Adults - non-fiction - Dewey number between 0-200 and 300-399 | 15. Adults - non-fiction - Dewey number between 400-499 and 500-599<br>16. Adults - fiction and poems<br>17. Adults - non-fiction - Dewey number between 700-799<br>18. Adults - Travel guides and audio books<br>19. Adults - non-fiction - Dewey number between 800-900 | 20. Music (CD/LP/Music DVD/Books/sheet music)<br>21. DVD/Blu-Ray<br>22. DVD/Blu-Ray<br>23. Share box with no. 24<br>24. Week loans / main storage in basement / christmas / easter / collections for schools<br>25. Cartoons<br>26. Dump / overflow / DVD for adults and children |

Figure 44: Setup in the sorting machine focusing on detailed sorting of children's books.

The bin layout presented in Figure 44 has merged bins that historically had low fill rates to make room for more bins for the children's books. The bins with changed sorting compared to the current setup are marked in purple in Figure 45. For example, non-fiction books with Dewey numbers 0-200 and 300-399 have been merged, and the bin for poems, audio books, and magazines has been distributed into the bins for fiction books and travel guides. In the new layout, some bins also have a new location in the sorting machine. For example, books going to other libraries in Norway are sent to bin 3 instead of bin 2 and reservations to be picked up at the main library are now sent to bin 11 instead of bin 9 in the sorting machine. The children's books are designated to bins 1, 2, 6, 7, 8, 9, and 10, marked in green in Figure 45.

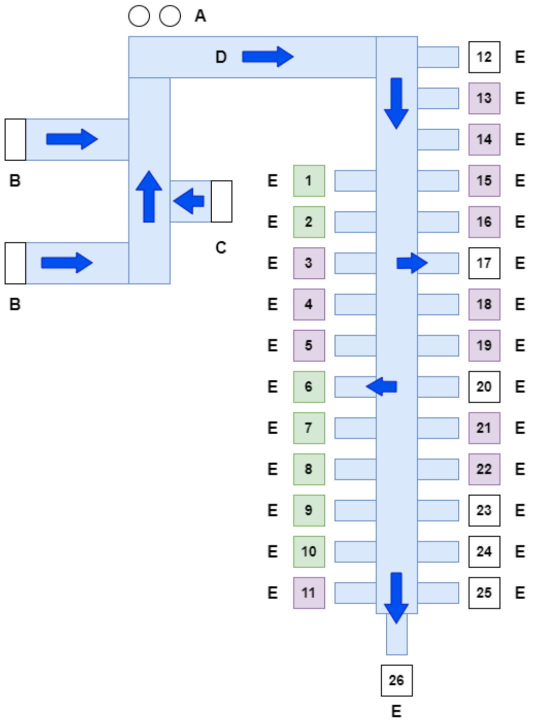


Figure 45: Bin setup in the sorting machine focusing on detailed sorting of children's books.

With this bin allocation, TPL has devoted seven bins to the sorting of children's books. Implementing these changes to the sorting machine resulted in a decrease in time spent on placing books on temporary and final shelves by librarians in the children and youth area. It became apparent that the temporary shelves were redundant, as the books could be sorted directly from the trolleys exiting the sorting machine into the final shelves. As a result, time spent on material handling was observed to decrease by an hour.

# 5 Predictive Analytics in a Library

This section continues to build on the results presented in the case study in Section 4, and aims at introducing predictive analytics in the library setting to answer RQ3 and RQ4. Firstly, operational areas that might benefit from predictive analytics will be presented, before it is examined how a dynamic sorting approach based on historical data for three different days would contribute to minimizing the time spent on material handling in TPL. This experiment can be compared to a perfect prediction model, as it is based upon having complete information about the incoming book flow. Subsequently, results attained by machine learning models trained to predict loan lengths, trained on the dataset containing loan and return transactions generated by Bibliofil as introduced in Section 3.2.1, are presented. Lastly, the data quality of this dataset will be reviewed.

## 5.1 Operational Areas Benefiting from Predictive Analytics

From the results presented in Section 4, capacity planning was observed to currently follow a predetermined, static, staffing list. Furthermore, there was identified several areas of improvement with regard to the sorting performed by the sorting machine. These results indicate that both capacity planning and the sorting machine are areas that could potentially benefit from utilizing predictive analytics to obtain predictions of future incoming book flow. Prediction of incoming book flow would provide information on multiple levels, based on the degree of detail:

- The number of incoming books each day.
- Which books will arrive at the main library, and therefore the sorting machine, each day.

Firstly, having knowledge of the incoming book flow can provide valuable insights for capacity planning, as it has been observed to be correlated with the time spent on material handling activities in the library. Additionally, by obtaining predictions regarding the specific books that will arrive at the main library, TPL would gain additional information about the material handling tasks related to holding and shelving, considering that different books necessitate different material handling tasks. Consequently, TPL would be able to make a more accurate estimation of the time needed for manual material handling tasks, thereby improving capacity planning efforts.

In the preceding chapter, multiple modifications implemented in the sorting machine displayed potential related to decreasing the time spent on material handling. Specifically, it appeared advantageous to enhance the sorting capabilities for children's books and allocate more bins to different branches. However, the current sorting machine is equipped with 26 bins, which means that if both the second and third proposals outlined in Section 4.4 were implemented, only nine bins would remain for sorting other books. This could decrease the level of detail in the sorting of these books, thus potentially yielding additional work when temporary shelving these books. Consequently, it is intriguing to explore the concept of dynamic bin allocation, which can adapt to the daily sorting demand.

By employing a dynamic system, the utilization rate of bins in the sorting machine can be increased, thereby minimizing wasted space by combining low-utilization bins. As stated in Bakkali et al. (2013), a general guideline is to employ a more dedicated or static storage approach when the input flow to the system is predictable, while a dynamic approach is preferable for unpredictable input flows. The advantage of a dynamic system based on sorting demand is visualized in Figure 46. However, for a dynamic system to operate optimally, TPL would rely on precise predictions regarding the upcoming book flow.

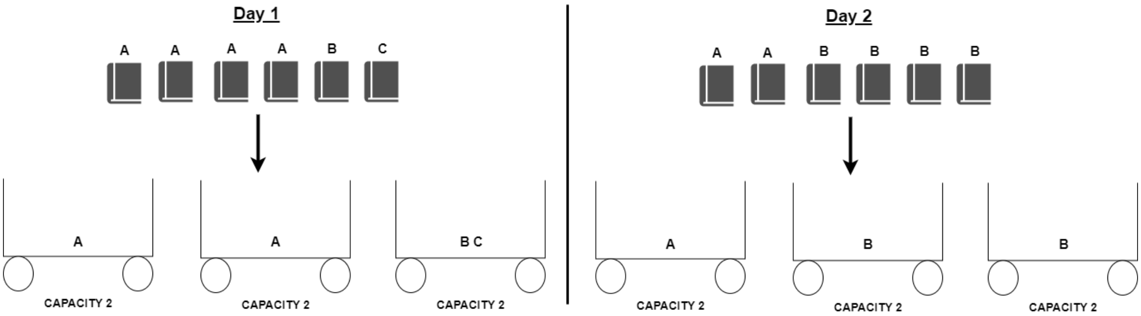


Figure 46: Benefits from a dynamic sorting system.

## 5.2 Potential of Dynamic Sorting

To demonstrate the potential benefits of implementing a dynamic sorting system at TPL, an analysis was conducted focusing on three days with the highest variations within the sample period: 26.01.23, 26.01.23, and 01.02.23. By examining the sorting logs from these days, the optimal sorting configurations were identified, akin to a scenario where a "perfect" prediction model is in place. The distribution of books in the different bins over the three days was:

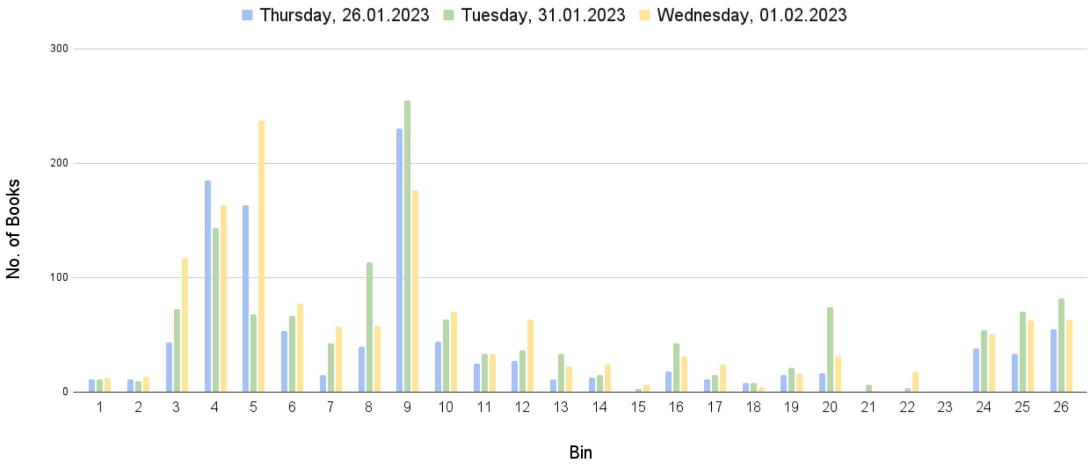


Figure 47: Bin distribution for 26.01, 31.01 and 01.02, where dynamic sorting could be applied.

From these statistics, it becomes apparent that, for instance, no books were to be sent to the prison branch on the 26th of January and that there was a spike in books that were to be sent to other branches on the 1st of February. To determine the branches associated with the books in bins 4 and 5, the sorting logs and data retrieved from Bibliofil had to be analyzed. The AQ code, described in Section 3.1.2, from the RX transactions in the sorting logs, indicates which branch the books in bin 4 are to be transported to, while statistics related to expedited reservations at the main library were used to determine where the books in bin 5 were to be transported. The distribution of books sent to different branches was as follows:

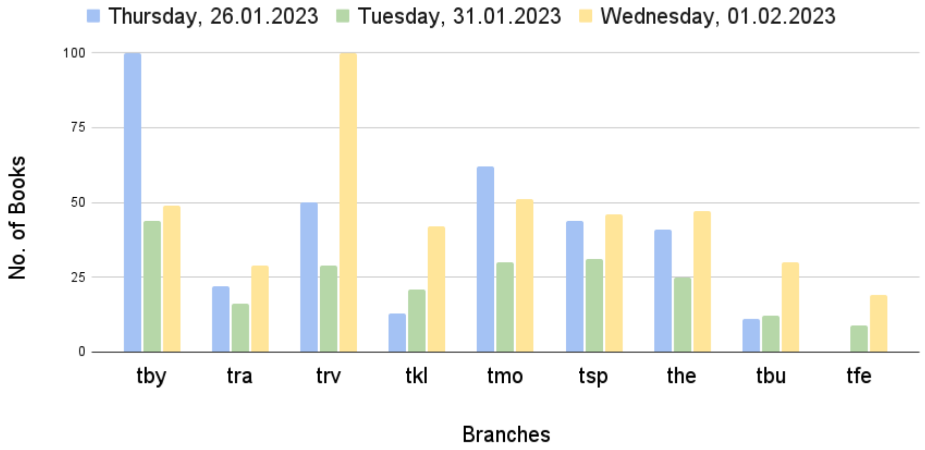


Figure 48: Statistics on the number of books being sent to other branches from the main branch on 26.01, 31.01 and 01.02.

The second area of interest was the incoming children's books, which fluctuated as depicted in the Figure 49:

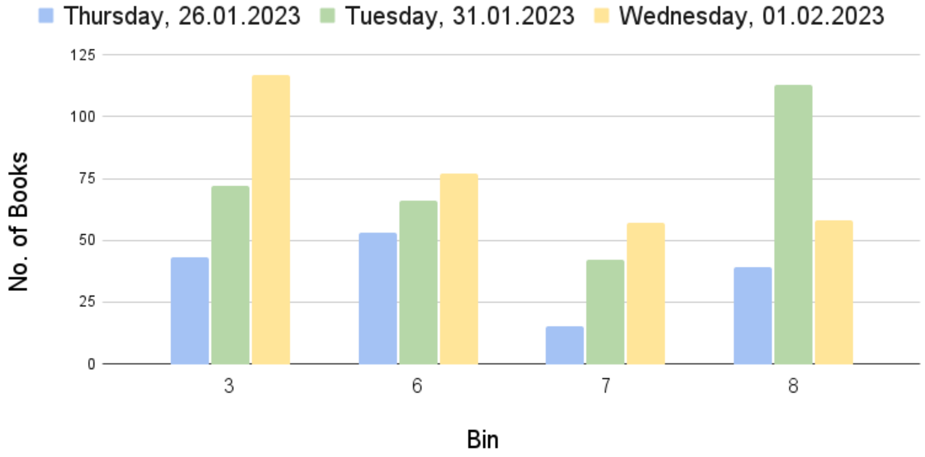


Figure 49: Statistics on the number of children's books coming in on 26.01, 31.01 and 01.02.

As mentioned earlier, it was identified that aligning the sorting process with the temporary shelves in the children and youth area would be beneficial. However, the distribution of books observed in Figure 49 does not indicate the specific temporary shelf where each book should be placed. The available data in the sorting logs do not provide sufficient detail to derive this information either. Therefore, it was determined that a static sorting approach would be employed for these books, allocating seven bins exclusively for children's books on each of the examined days, in order to guarantee no time was spent on temporary shelving in the children and youth area.

The following paragraphs will present the setups derived for each of the three days.

### Setup 26.01

The incoming book flow on the 26th of January can be described by having a relatively high proportion of adult books, whereas no books were sorted in bin 15 and the prison branch (tfe). Besides prioritizing the seven bins for children's books and books going to other branches, the bins with no books allowed reasonably detailed sorting among the adult books. The derived bin setup is presented below:

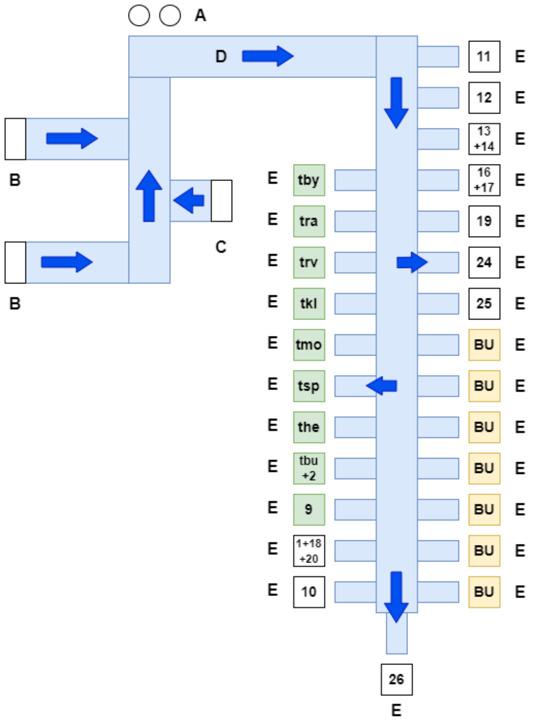


Figure 50: Proposed bin setup for 26.01

The proposed setup required 27 bins containing adult books to be emptied, compared to 29 bins with the AS-IS setup, which explains the minor decrease observed in task 6. Furthermore, the number of books that required scanning decreased from 359 to 27. Sorting in temporary shelves in the children and youth area could be neglected. To summarize, time spent on manual material handling tasks would have changed:

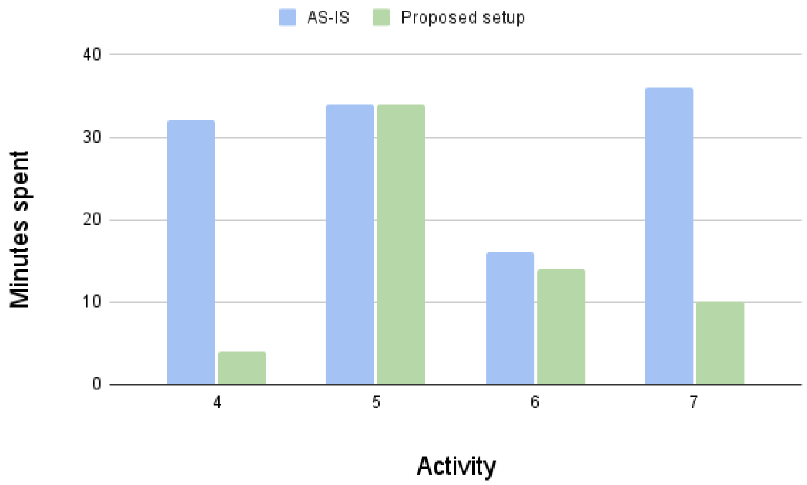


Figure 51: Comparison of the amount of manual material handling with the current setup and the proposed setup on the 26th of January.

In total, 56 minutes would have been saved by applying this setup on the 26th of January, compared to the AS-IS setup. The majority of the decrease in time spent stemmed from the drastic decrease in time spent on prepping books for further transport and eliminating temporary sorting in the children and youth area.

### Setup 31.01

Continuing with the 31st of January, both the prison bin and bin 15 had to be considered, which forced changes to be made in the setup proposed for the 26th of January. The derived bin setup for the 31st of January was as follows:

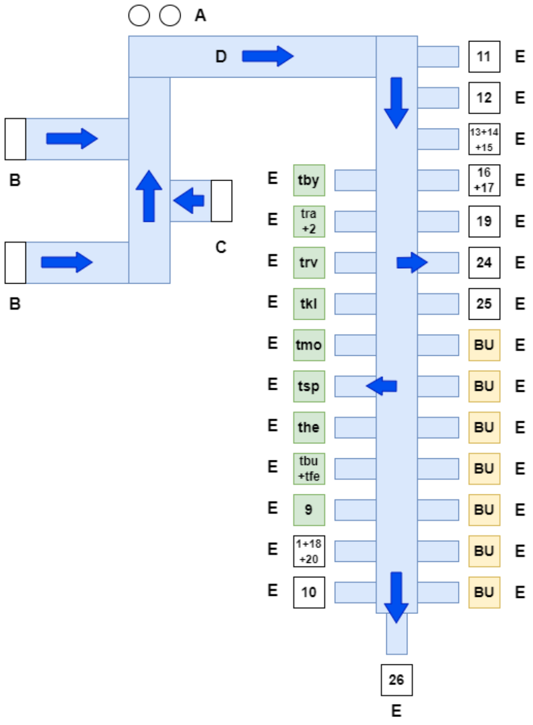


Figure 52: Proposed bin setup for 31.01

With the proposed setup, a total of 30 bins containing adult books had to be emptied, compared to 32 bins with the AS-IS setup, explaining the decrease in minutes spent on task 6. Furthermore, the number of books having to be scanned or manually sorted as they were prepared for transport for other branches was cut from 229 to 49. Sorting in temporary shelves in the children and youth area could be neglected. To summarize, time spent on manual material handling tasks would have changed as depicted:

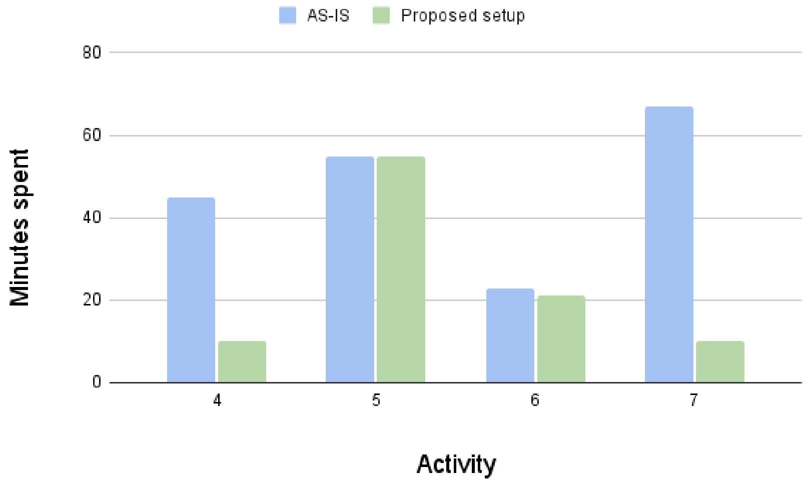


Figure 53: Comparison of the amount of manual material handling with the current setup and the proposed setup on the 31st of January.

In total, time spent on manual material handling tasks was decreased by 94 minutes, which equates to approximately a 16 percent decrease.

### Setup 01.02

The following day, the 1st of February, was characterized by a large increase in books going to other branches, resulting in it being beneficial to devote one bin to each branch. Consequently, the following setup was derived:

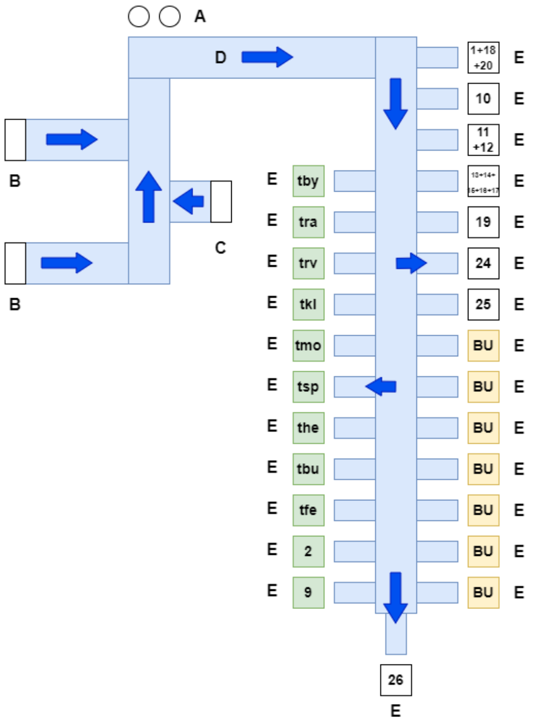


Figure 54: Proposed bin setup for 01.02

With the proposed setup, a total of 27 bins containing adult books had to be emptied, compared to 31 bins with the AS-IS setup. However, based on the decrease in the detail of sorting of these books, time spent on temporary shelving these books, task 6, was unaffected. The proposed setup saved 106 minutes compared to the AS-IS. With this setup, the material handling related to preparing reservations for other branches was eliminated as each branch had its own bin.

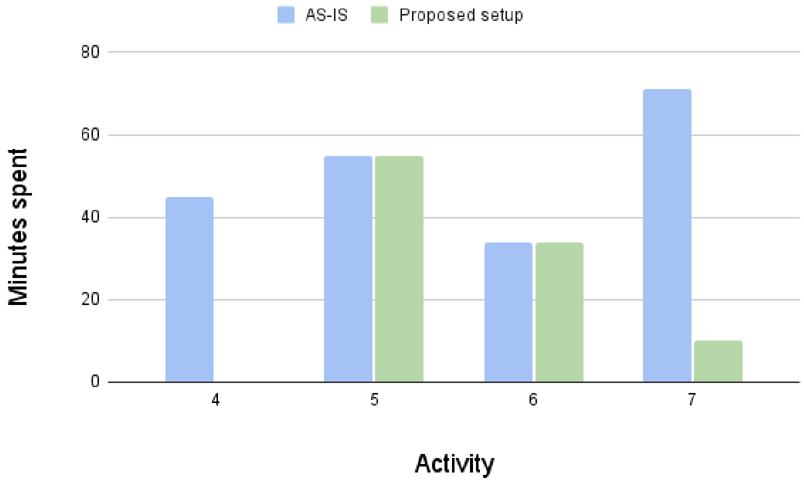


Figure 55: Comparison of the amount of manual material handling with the current setup and the new proposed setup for the 1st of February.

### Comparison of the Different Setups

Figure 56 summarizes the decrease in time spent on material handling tasks with the setups employed on the three different days. As illustrated, most of the time saved was related to the tasks of handling reservations to other branches and placing books on temporary shelves in the children's department. There were also minor reductions related to moving books to temporary shelves in the adult department caused by fewer full bins, on two of the days.

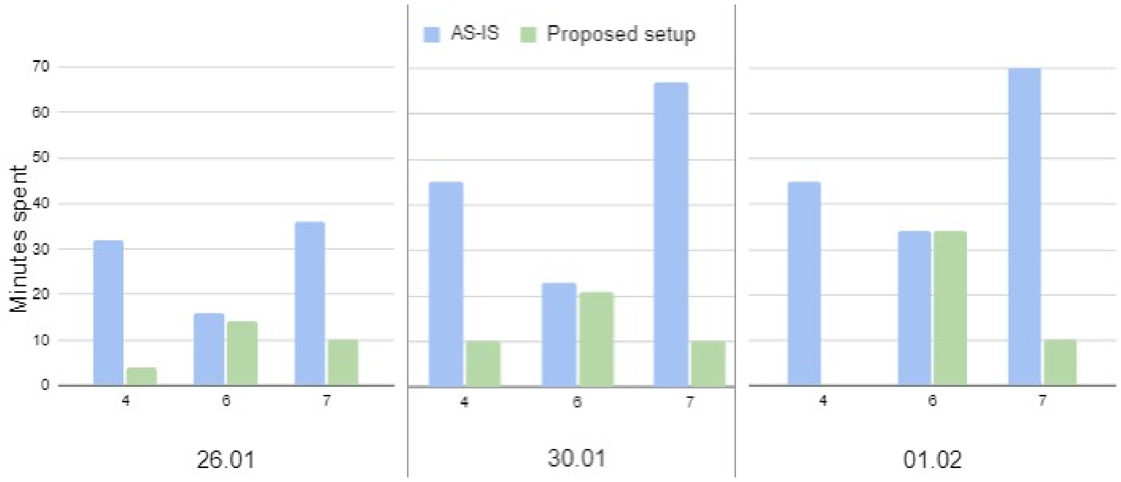


Figure 56: Summary of changes in time spent on material handling tasks with dynamic sorting.

However, to compare the performance of the different setups, it is necessary to investigate how the proposed setups would perform on each of the different days. The table below presents time spent on material handling with each of the setups for the three days:

| | 26.01.23 | 31.01.23 | 01.02.23 |
|---|---|---|---|
| Setup 1 | 395 | *(491)* | *(499)* |
| Setup 2 | 398 | 491 | 499 |
| Setup 3 | 412 | 497 | 487 |

Table 10: Comparison of manual work in minutes with the different setups and input flows

The number of minutes spent on manual material handling tasks by using setup 1 on the 31st and 1st is set to be equal to that of setup 2. This is because when the prison and bin 15 are included in setup 1, it becomes nearly identical to setup 2. Although the differences in time saved on material handling between the three setups are not substantial, it is intriguing to observe that the setups outperform each other on different days when considering the implementation of dynamic bin allocation.

## 5.3 Applying Machine Learning Models to Predict which Books that are Returned Each Day

After examining the potential yielded by what can be considered a perfect prediction, an obvious continuation is to examine the feasibility of actually predicting future incoming book flow. Consequently, this section presents the results obtained by the machine learning models, explained in Section 3. Firstly, the performance of the initial models is presented, before the results of applying ensemble models are presented. Finally, the performance of these machine learning models is compared with the performance observed with utilizing constant prediction models.

### 5.3.1 Performance of Standard Models

Overall, the models failed to predict loan lengths with acceptable accuracy. Three different models were applied to the problem. Firstly a simple mode based on linear regression was applied. Further, a decision tree was applied. Finally, a more complex model was introduced, namely an Artificial Neural Network in the form of a Multi-Layer Perceptron, which is a regular Feed Forward Artificial Neural Network. Hyperparameters for the decision tree and MLP were determined by a grid search. For the decision tree, the maximum depth was set to 13, and the minimum samples in a leaf were set to 2. For the MLP, alpha was set to 0.001, batch size was set equal to 10, and with hidden layers of size 10 and 5. Maximum iterations were set to 5000.

| **Model / Metric** | Linear Regression | Decision Tree | Multi-Layer Perceptron |
|---|---|---|---|
| MAE | 11.501 | 10.420 | 12.668 |
| MAPE | 1.251 | 1.062 | 1.367 |
| RMSE | 14.178 | 12.291 | 15.209 |
| $R^2$ | 0.131 | 0.236 | 0.000 |

Table 11: Results from the first machine learning models

Furthermore, an SVR was also trained on the dataset, but was not able to produce predictions. The reason for this may have been caused by the size of the dataset. The fit time complexity of SVRs is over quadratic which can make it hard for the model to fit on bigger datasets (Developers 2023).

The MAE values indicate that none of the models were able to generate predictions with a lower average deviation than 10.42 days from the correct loan length. Prediction errors of that magnitude do not offer significant business value to the library. In the case of these models, the RMSE yields the same ranking amongst the models. The same applies to the MAPE observed in the predictions. When interpreting the MAPE as a percentage, it needs to be multiplied by 100. Based on the MAPE, it is apparent that the average prediction error of any of the models is in excess of 100%. However, given that the average loan length is 25 days and the MAE is between 10 and 13, the percentage is not a precise estimation of the actual prediction error. The MAPE rather explains that the models struggle to predict points far away from the mean. The $R^2$ for all the models can be considered to be low. As mentioned in Section 2.3.6, a low $R^2$ indicates that the variation of the dependent variable, loan length, is poorly explained by the independent variables. Consequently, any correct predictions may be attributed to chance rather than a meaningful relationship between the variables. Further, by comparing the training results with the test results, it was evident that the DT model was overfitting its predictions.

### 5.3.2 Performance of Ensemble Models

As a means of trying to improve the prediction results, some ensemble models were applied to the prediction problem. The ensemble models Random Forest, Adaptive Boosting, and Gradient Boosted Decision Trees, were evaluated due to the DTs propensity to overfit, as these models are well-known for their ability to handle overfitting better than individual decision trees. The results indicated that both the RF and GBDT outperformed the DT. Hyperparameters were once again determined by grid search. The hyperparameters for the best performing model, the RF, were as follows: minimum sampler in a leaf was set to 3, the number of trees was set to 2000, and the random state was equal to 42. However, regardless of the improvement, the prediction accuracy is still too poor in order for the library to make any business value the predicted outcomes. Moreover, both RF and GBDT also showed tendencies of overfitting their predictions.

| **Metric \ Model** | Random Forest | Adaptive Boosting | Gradient Boosted Decision Trees |
|---|---|---|---|
| MAE | 7.295 | 11.591 | 8.698 |
| MAPE | 0.752 | 1.267 | 0.874 |
| RMSE | 10.372 | 14.254 | 11.427 |
| $R^2$ | 0.535 | 0.122 | 0.436 |

Table 12: Results from the ensemble machine learning models

## 5.4 Constant Models

Due to the machine learning models' lack of accuracy, the performance of simpler constant models was examined, to further assess the machine learning models' performances. The constant models tested assume that each loan was returned X days after the day the book was loaned. X was chosen based on the presence of the different loan lengths in the dataset, as depicted below:

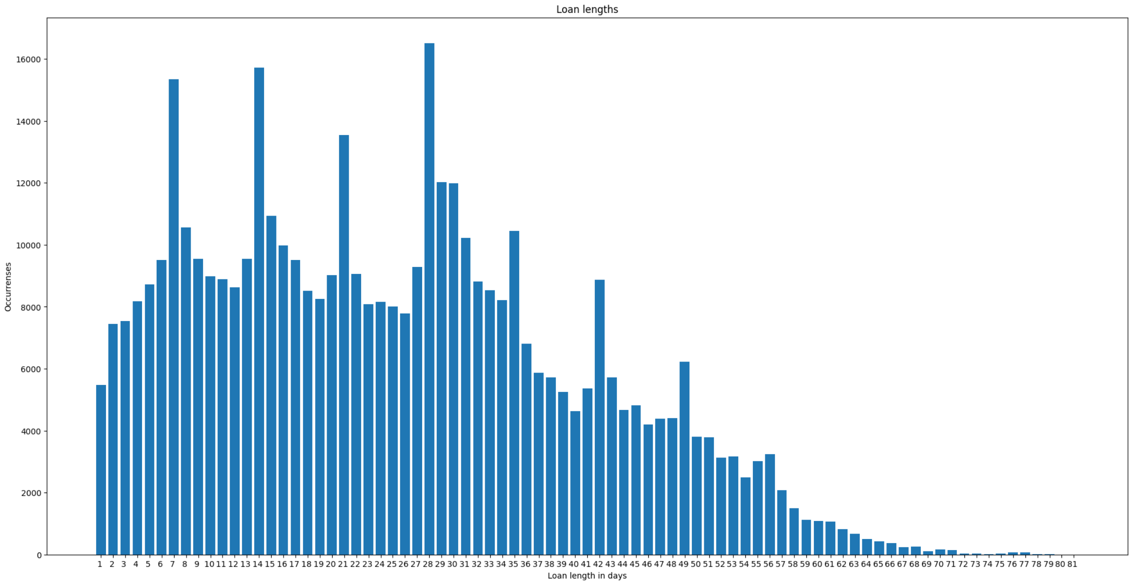
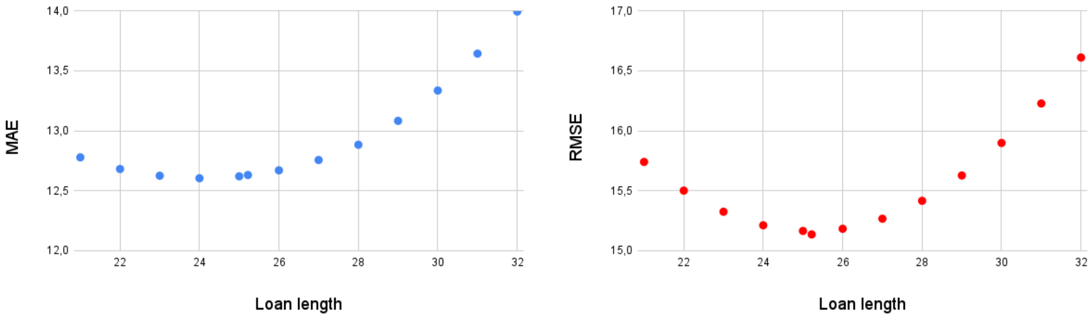


Figure 57: Distribution of the length of the loans in the transaction dataset.

The average loan period was observed to be approximately 25 days, the median was 24 days, and the most common loan length was 28 days. The performance of the constant models with X around this range was reviewed. In Figure 58a and Figure 58b, it becomes clear that the global minimum for the constant models stems from an X of 24 and approximately 25 days, considering both MAE and RMSE. This corresponds well to the median and average loan length.



(a) Distribution of MAE for loan lengths between 21 and 32 days

(b) Distribution of RMSE for loan lengths between 21 and 32 days

Figure 58: Distribution of MAE and RMSE for constant models for loan lengths between 21 and 32 days

The performance of the constant model with the optimal X-values is presented in the table below:

| **X value / Metric** | 24 | 25.22 |
|---|---|---|
| MAE | 12.604 | 12.631 |
| MAPE | 1.300 | 1.368 |
| RMSE | 15.212 | 15.136 |
| $R^2$ | -0.006 | 0.000 |

Table 13: Results obtained by the constant models.

To summarize, the constant models exhibit similar performance to the machine learning models, excluding RF and GBDT, apart from the $R^2$ metric. The $R^2$-value being close to zero, or zero, for the constant models can be explained by the independent variable, X, being static, thus not explaining any variation in the loan lengths.

## 5.5 Data Quality

Given the lackluster results obtained from the prediction models, it is interesting to assess the data quality of the dataset generated by the library. As mentioned in the Methodology chapter, the original dataset had to be modified in order to provide information about loan lengths. When necessary, the dimensions describing the quality of the dataset have been assessed after this modification had been made.

### Completeness

As mentioned in Section 3, at first glance, all transactions in the dataset were complete, as missing values, or null values, were not present. However, after further inspecting the dataset, it was observed that some transactions were assigned the value “unknown” for some of their variables. Consequently, this affected the completeness of the data. In total, 127 678 out of the 470 000 transactions contained an unknown value, which equates to a completeness of 0.728, when using the metric proposed in Section 2.4. Different preprocessing techniques, described in Section 3.2.2, were applied to increase completeness.

### Accuracy

Most of the variables in the dataset are static values, retrieved from Bibliofil. There is almost impossible to verify most of these values. However, some extreme or obscure values were observed; 70 132 transactions had an age value of either a negative number or a number over 102, and some had invalid postal codes. This is likely a result of faulty data in the user catalog. The other variables, such as timestamps for loan and delivery, are generated by the RFID-system, which seems reliable and accurate based on the examination of the dataset.

### Consistency

Consistency is normally used to evaluate how records from different datasets compare. In this case, this can be compared to the loan- and return transactions. Normally, if the dataset was consistent, it would likely have the same number of loans and returns. However, this was not the case. In fact, it was approximately five times more return transactions compared to loan transactions. Moreover, there were some discrepancies in the timestamps present in the dataset, and the timestamps in the sorting logs. However, this was not seen to impact the quality of the dataset for this purpose, as the dates matched.

### Timeliness

The transaction dataset involves data that is continuously generated in real-time. Therefore, it can be said to be of acceptable timeliness for the purpose of predicting the length of loans.

### Fitness for Purpose

The fitness for purpose is likely the main demise of this dataset in the context of predictive analytics. As mentioned, the dataset consists of mainly static variables and general information about the loaner. There was observed little to no relationship between these variables and the target variable, loan length, as illustrated in the correlation matrix.

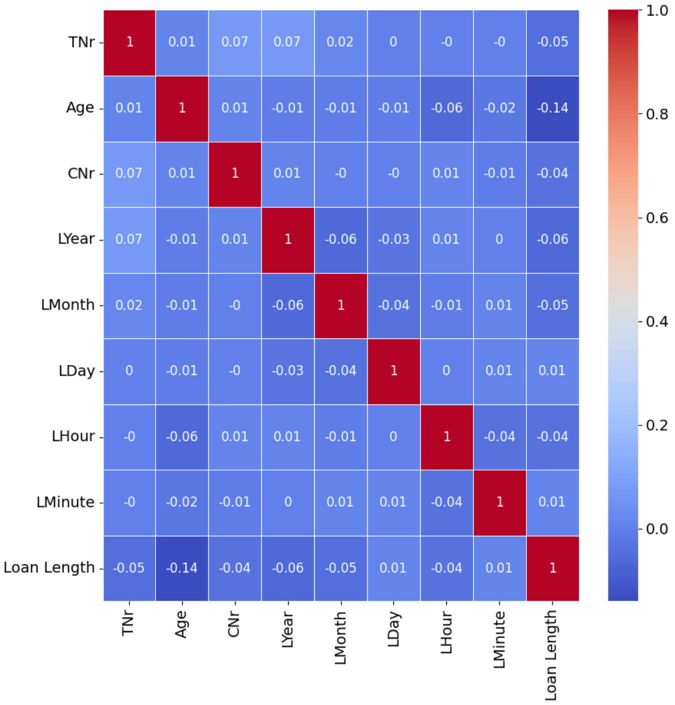


Figure 59: Correlation analysis chart of the numerical values in the dataset.

The feature with the highest correlation with loan length was observed to be the age variable, but also this correlation was close to non-existing. Usually, features that correlate poorly with the target variable should be classified as irrelevant and removed. However, as mentioned in Section 3.2.2, this was not possible for this dataset, as it was not impossible to derive a valid threshold for what should be kept or removed, given that all variables portrayed similar, low, correlation.

# 6 Discussion

This thesis has been written as a contribution to the SmartLIB project, a cooperation project between Trondheim Public Library and NTNU. The main goal of the project is to explore possible technological implementations at TPL, to reduce manual material handling, thus allowing librarians to devote more time to service patrons and other value-adding activities. In particular, this thesis has explored how material handling of incoming books is currently conducted, and subsequently how much time is required to manage the incoming book flow. Further, it has been examined how different operational areas could benefit from predictive analytics. This section aims to discuss the findings from Section 4 and Section 5, and answer the four RQs presented in the Section 1:

- RQ1: How does incoming book flow affect material handling in a library?
- RQ2: How can a library enable more efficient material handling?
- RQ3: How can predictive analytics be applied in a library setting to enhance the planning of material handling?
- RQ4: How can a library enhance data quality to make predictive analytics more applicable?

These RQs are discussed in chronological order, following a similar structure as to how the results were presented. An understanding of the previous RQs is beneficial for answering subsequent RQs, as depicted in Figure 3.

## 6.1 Relationship Between Incoming Book Flow and Material Handling in a Library

In Section 4.2, the focus was on analyzing the material handling tasks involved in managing the incoming books, which can be considered a significant aspect of material handling in a library. The process of managing the incoming material flow was divided into four distinct steps: material return, sorting, holding, and final shelving. Throughout these steps, a total of ten material handling tasks were identified:

| Number | Material Handling Tasks |
|---|---|
| **Step 1** | All manual work related to non-collected reservations. |
| | All manual work related to returns in the outdoor boxes. |
| | All manual work related to getting full boxes from the garage and returning the books to the sorting machine. |
| **Step 2** | Automated sorting in the sorting machine |
| **Step 3** | Reservations to be sent to other branches in the city and other Norwegian libraries. |
| | Reservations to be collected at the Main branch incl. the time it takes to put in on the reservation shelf. |
| | Moving full trolleys to the temporary shelf at the adult department at the main branch. |
| | Moving full trolleys to the temporary shelf at the youth and child department at the main branch. |
| **Step 4** | Moving books to the original shelf |
| | Moving books to the basement |

Table 14: Overview of the steps and related material handling tasks.

The time staff spent on material handling and the volume of the incoming book flow in the examined period developed as depicted in the figure below:

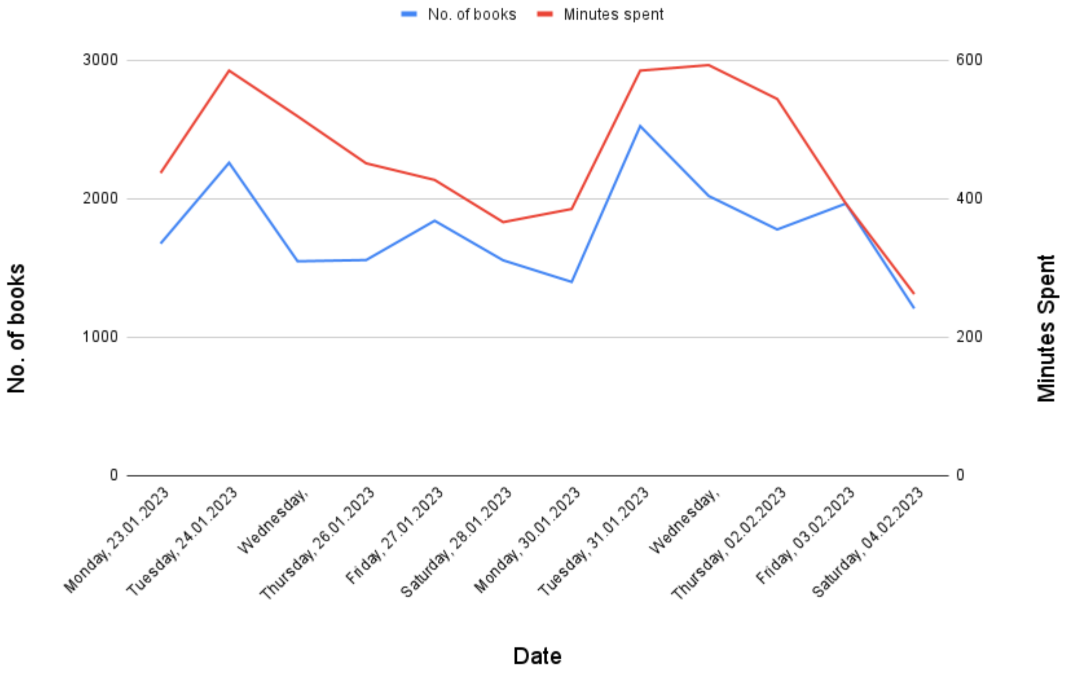


Figure 60: Books delivered and prepared for reservation plotted with minutes spent on manual work.

Results indicated a strong correlation, with a correlation coefficient equal to 0.762, between the time spent on material handling tasks amongst employees and the volume of incoming book flow. In the subsequent paragraphs, it will be discussed how the material handling tasks related to the different process steps are affected by the incoming book flow.

## Material Return

Figure 61 depicts how the time spent on material handling tasks categorized as step 1, material return, fluctuated in comparison to the incoming book flow.

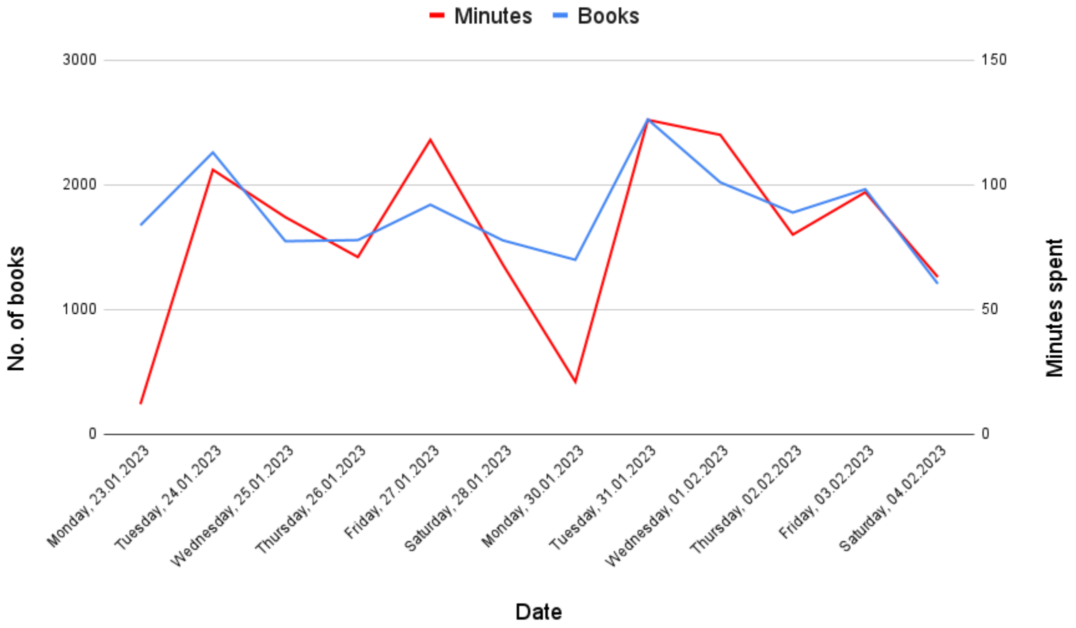


Figure 61: Books delivered and prepared for reservation plotted with minutes spent on material handling tasks categorized as material return.

The time spent on material handling tasks related to step 1 displayed a correlation coefficient of 0.695 with the volume of the incoming book flow. This strong correlation can best be explained by the time spent on feeding books into the sorting machine, which increases linearly with the number of incoming books. However, it should be considered that the number of books fed into the sorting machine by library staff, does not perfectly resemble the volume of the incoming book flow, as a portion of the books are returned by patrons. Generally, however, a larger book flow should result in more books being fed into the sorting machine by staff, thereby explaining the relatively high correlation.

## Sorting

Sorting at the main library is automated by the sorting machine, meaning that no time is spent on sorting by library staff. Therefore, the time spent on sorting in the library is not considered to be affected by the incoming book flow. However, as previously mentioned, downtime in the sorting machine can occur, which results in sorting having to be performed by librarians. In this case, the time spent on sorting would probably be greatly affected by the volume of the incoming book flow.

## Holding

Furthermore, results indicate that the holding tasks in the library are somewhat affected by the incoming book flow. The time spent on these material handling tasks displayed a relationship with the volume of the incoming book flow, with a correlation coefficient of 0.537. The relationship is further illustrated in the following graph:

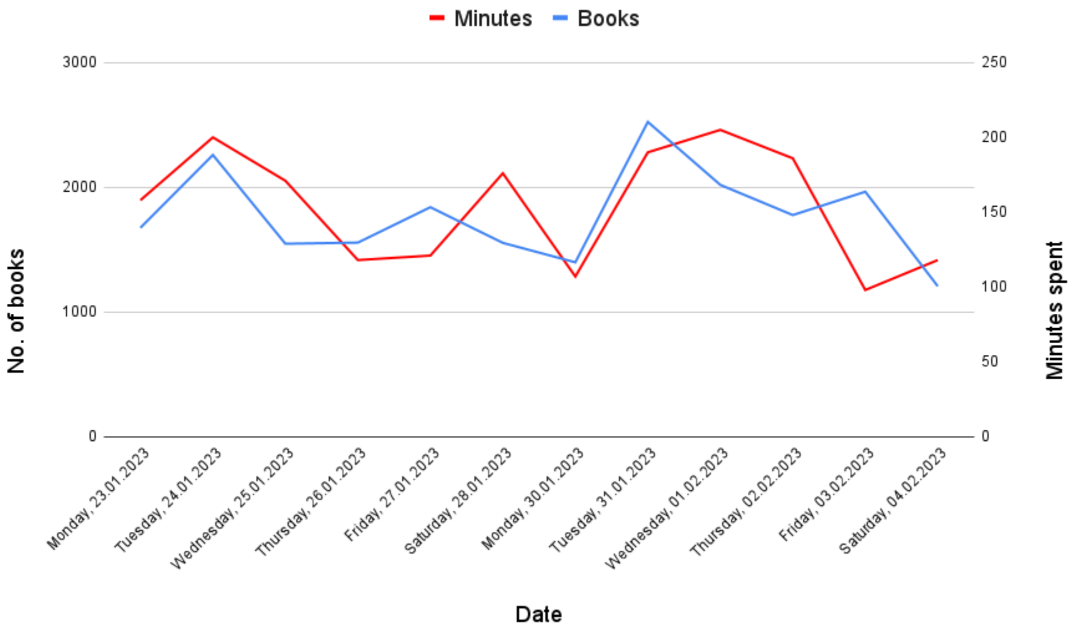


Figure 62: Books delivered and prepared for reservation plotted with minutes spent on material handling tasks categorized as holding.

Consequently, it appears like these tasks are not solely dependent on the volume of the incoming book flow, but also the content of it. For instance, holding tasks will require more time if a lot of children's books arrive in the system, as the temporary shelving of these books is cumbersome in the current system. The same applies for the time spent on preparing reservations or transport to other branches. In fact, by analyzing Figure 62, it is possible to calculate that it requires, on average, approximately 7.7 seconds of material handling to perform the holding steps on one book. However, by examining the time spent on temporary shelving of children's books, it is apparent that these books require on average 14.4 seconds of material handling related to holding. These discrepancies further highlight that time spent on material handling related to holding is dependent on which books constitute the incoming book flow. However, a larger book flow will generally increase the time spent on these tasks as more books require material handling, and the likelihood of more children's books arriving in the library increases.

## Final Shelving

Lastly, results indicated that final shelving was the process step least dependent on incoming book flow, with an observed correlation of 0.495.

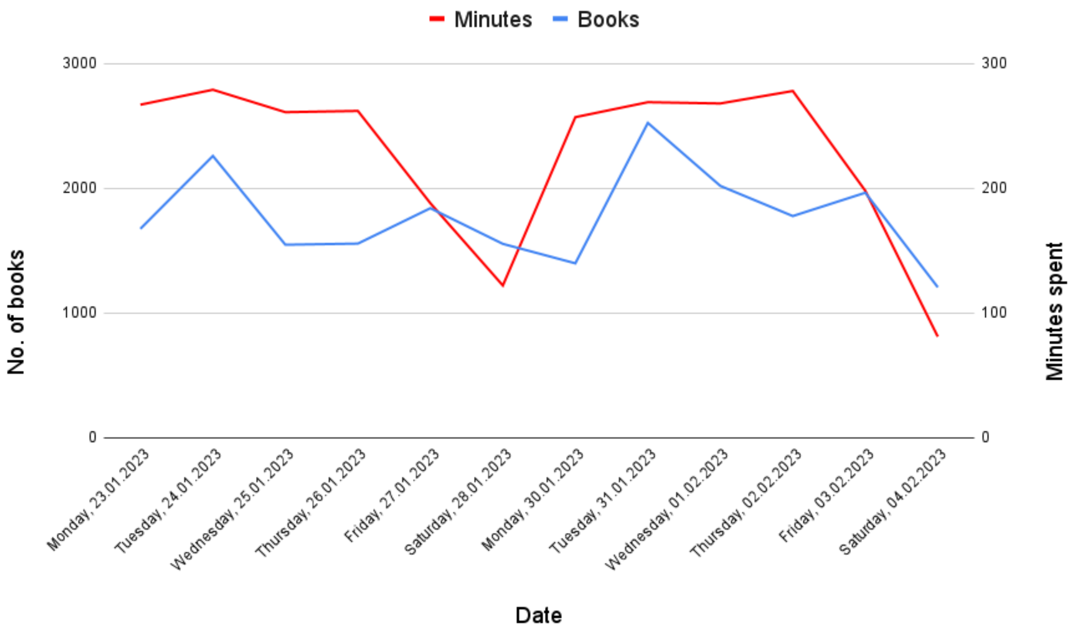


Figure 63: Books delivered and prepared for reservation plotted with minutes spent on material handling tasks categorized as final shelving.

This seems to contradict the results presented in the previous paragraphs. Logically, time spent on final shelving should be affected similarly by the incoming book flow as the previously performed material handling tasks. However, the number of books necessitating final shelving each day in the sample period displayed a correlation coefficient of only 0.71 with the total number of incoming books each day, which partly explains the decrease in correlation. Furthermore, there are likely two other factors that have contributed to the low correlation. Firstly, as explained in the in Section 4, the time spent on final shelving is characterized by large fluctuations, determined by other factors than the volume of the incoming book flow. For instance, librarians might service patrons simultaneously as performing final shelving, which greatly affects the time spent on this task. Secondly, the results are based on daily correlation, which might not be an adequate way of describing the relationship between incoming book flow and time spent on final shelving, given how the library operates. As mentioned in Section 4.2.4, the temporary shelves on the second and third floors are consciously not emptied each day, as they are used as sort of “recommendation shelves”. Consequently, there might be a lag related to how the volume of incoming book flow affects the time spent on this task, thus making the relationship difficult to precisely describe.

### Summary

The results from Section 4 provide new insight into the relationship between incoming book flow and material handling tasks in a library. Overall, time spent on material handling was observed to strongly correlate with the incoming book flow, summarized in the table below:

| **Description** | **Correlation** |
|---|---|
| Material Return: | 0.695 |
| Sorting: | —- |
| Holding: | 0.535 |
| Final Shelving: | 0.495 |
| Total: | 0.762 |

Table 15: Summary of the material handling steps related to incoming book flow.

It should be considered, however, that correlation does not necessarily imply causation. Nonetheless, the strong relationship supports the general notion that it demands more time to process an increase in input. Moreover, a larger input flow will no matter how operations are conducted, result in more books going through the four process steps mentioned above, thus needing material handling. Consequently, when performing certain material handling tasks, librarians will either have to perform heavier lifting or execute tasks multiple times. Considering the discomfort of performing heavy and awkward lifts, it can be assumed that librarians aware of ergonomics choose to instead perform more lifts, thus increasing the time spent on material handling.

Finally, it is important to acknowledge that the sample size of twelve days examined in this study is relatively limited in duration. Consequently, drawing definitive conclusions regarding the observed patterns or trends becomes challenging, as the potential influence of coincidental factors on the results cannot be entirely discounted. Moreover, the length of the sample period further introduces an element of uncertainty in the recorded durations of the material handling tasks. The recorded durations of the material handling tasks are likely to possess a degree of imprecision, primarily due to their manual recording and the nature of certain tasks being performed in a continuous or loosely defined manner, such as scanning reservations. The observations somewhat deviated from the perception of the employees, as they usually estimated that more time was spent on material handling tasks. A plausible explanation for this is that waiting times might occur sporadically when librarians are performing certain material handling tasks, resulting in the librarians being occupied for longer than the time it actually requires to perform a task. Additionally, another possible explanation of the perceived lengthy durations of material handling tasks can be attributed to librarians finding such tasks monotonous and unmotivating, thus overestimating the actual time invested in performing these tasks. Nevertheless, to enhance the robustness of the findings, the inclusion of additional quantitative data would have been advantageous, as it would have augmented the overall evidentiary strength of the study.

## 6.2 Increasing the Efficiency of Material Handling in a Library Setting

It is apparent that a considerable amount of time is allocated to material handling at TPL, which is likely generalizable to other libraries, with a similar degree of automation in the material handling system. In the twelve-day period examined, it was uncovered that around 5550 minutes were spent on manual material handling in order to manage the incoming book flow, which equates to 462.25 minutes daily. It is of great interest to decrease this time, in order for librarians to be able to devote more time to servicing patrons, thereby enhancing the quality of library services. There are several possible ways for a library to achieve more efficient material handling. The subsequent paragraphs will discuss some of these, in light of how material handling is conducted at TPL.

### Automation

In TPL's operations, it is evident that the sorting machine massively decreases the time spent on material handling by staff. Currently, the sorting machine is the only robotic solution integrated into the material handling system. It is obvious that with the implementation of more robotics, the time spent on material handling by staff will further decrease, as a larger part of the process is automated. An example of robotic technologies that could increase the efficiency of material handling in a library is AGVs, similar to the ones employed by Oodi library, as mentioned in Section 2.1. These robots are capable of performing most of the necessary inter-library movements of books, with the exception of shelving tasks.

However, as already mentioned, integrating technology that enables automation is costly, and might be infeasible for several libraries, or not worth the investment. Furthermore, new technologies would likely be integrated incrementally, possibly with a large time horizon. Therefore, it is interesting to instead analyze the current operations at TPL, and consider potential improvements in the process of managing the incoming book flow, that could be implemented within the current system.

### Facility Layout

Disregarding the implementation of robotic solutions, streamlining material handling tasks associated with the management of the incoming book flow presents inherent challenges, at least in the case of TPL. However, one possibility is to analyze the facility layout. Certain tasks, such as retrieving books from the basement, are particularly complex to optimize without altering the library facilities or robotic automation. Currently, the process necessitates an employee going to the basement via an elevator and subsequently transporting the books on trolleys up to the first floor. Decreasing the time spent on this task is inherently difficult without repositioning the sorting area in closer proximity to the basement, which would demand huge investments.

Moreover, there also exist simpler changes to the library layout, that could increase the efficiency of material handling. Examples include positioning, for instance, temporary shelves and the sorting area, to optimize the movement of books, and eliminate waste present in the movement of books. However, there are several constraints that would have to be considered, and these smaller changes in the facility would probably have limited effects.

### Enhancing the Sorting

However, the results in Section 4 uncovered areas that could increase efficiency in material handling, without drastic changes in the facility or investments in new technology. An operational domain that was observed to have a significant impact on the duration of material handling in the library was the manner in which books were sorted within the sorting machine. The sorting machine has been highlighted as a vital component of TPL's material handling system, and subsequent material handling tasks were found to be reliant upon the sorting within the machine. Examples of such tasks include:

- Emptying bins around the sorting machine.
- Temporary shelving.
- Preparing books for reservations.
- Preparing books for transport to other branches/libraries.

Consequently, several alterations to the setup of the sorting machine were reviewed in Section 4, with the aim of lowering the time spent on material handling at TPL. The main focus areas were to level the fill rates in the sorting machine, and enable more detailed sorting for books transported to other branches and children's books.

Firstly, the effects of leveling the fill rates were examined. It was observed that this would not contribute to decreasing the time spent on material handling tasks at TPL, as the total number of full bins actually increased, although marginally, in the examined period when the fill rates were leveled. However, it does provide the opportunity to empty multiple trolleys simultaneously, given that the capacity of the expanded bins increases, which could decrease the number of trips from the sorting machine to temporary shelves. This was not an observed effect at TPL, likely explained by the fact that all the bins that were most frequently filled, and thus expanded, sorted books that shall either be prepared for reservation or transport to another branch, meaning that they do not have to be moved to temporary shelves. Leveling the fill rates would, however, ensure that the bins were filled more evenly throughout the day. Consequently, staff would have the possibility to perform the material handling in bulk, at a desired time, possibly alleviating pressure and stress on the librarians operating in the sorting area. As expressed by head of IT at TPL, it would be of great pleasure for the librarians to "have the sorting machine work after their tempo, and not the opposite".

Secondly, the current sorting system's deficiency in detail regarding books prepared for transport to other branches was identified. With two bins sorting books for eight branches, additional sorting became necessary to place the books correctly in designated transportation boxes, posing inefficiencies. Exploring alternate bin allocations, it was found that dedicating one bin to each branch could eliminate the need for additional sorting and save 30 minutes of daily material handling time, and it additionally leveled the fill rates in the sorting machine.

Lastly, the sorting of children's books emerged as a significant issue in the current operations, requiring substantial time for temporary shelving. To address this problem, adjustments were made to the sorting machine to align with the temporary shelves in the children and youth area, eliminating the need for temporary shelving in this area. By assigning seven bins to correspond with the seven temporary shelves, the time spent on this process step was eradicated, resulting in approximately an hour of daily material handling savings.

From these experiments, it became evident that implementing several adjustments in the sorting machine would yield advantageous outcomes, and it highlights the potential for optimizing sorting machine configurations to improve material handling efficiency in libraries. Nonetheless, the sorting setups were not implemented in practice, only simulated, thus adding a degree of uncertainty to the observed effects. To be able to confirm the effects, the sorting setups would have to be tested in practice.

### Summary

To summarize, three potential means for increasing the efficiency of material handling have been discussed. These are repeated in the table below, alongside how they would contribute to enhancing the material handling.

| | |
|---|---|
| Automation | Automate material handling by integrate robotic solutions in the material handling system. |
| Facility layout | Minimize waste related to the movement of books in the library. |
| Sorting | Eliminate unnecessary processes in the material handling. |

Table 16: How and why the efficiency of material handling in a library can be increased.

However, as discussed, there are several hindrances related to automation and facility layout. Consequently, this thesis mainly investigated how the sorting aspect could increase efficiency in the material handling processes, and the results from this thesis have provided new insight into how sorting in a library affects subsequent material handling tasks, by examining the effects of several alterations in the current system at TPL. It is imperative, however, to acknowledge that the suitability of the proposed alterations may vary across different libraries, as the challenges identified through this case study might be specific to TPL, thus limiting the generalizability of the results. Consequently, it is essential to carefully evaluate the impact of any changes and consider the unique context of each library's operations to ensure the best outcomes are achieved. Nonetheless, the essence of these findings does not lie in the specific changes that result in reduced material handling time, but rather in the benefits achieved through their implementation, namely the elimination of excessive process steps. Eliminating process steps of the material handling can also be considered to be beneficial from an ergonomics standpoint, as it contributes to decreasing the movement of books in the library. Moreover, awareness of ergonomics in the material handling system would decrease the likelihood of injuries and increase the well-being of employees, thereby likely increasing long-term efficiency.

## 6.3 Operational Areas Potentially Improved by Predictive Analytics

Statistics derived from the transaction dataset indicate that loan lengths usually differ from one to approximately 50 days, quite evenly distributed, disregarding the right tail of the distribution, as depicted in Figure 57. However, one pattern that became apparent from the statistics is that there are peaks for loan lengths of one, two, three, and four weeks. These results are supported by the results from T. H. Lee and J. W. Lee (2021), although the study focused on academic libraries. In this study, it was observed a weekly periodic pattern for deliveries before the deadline, while books delivered after the deadline was characterized by a more complex and less predictable pattern. From the statistics of loan lengths, most loans were observed to be returned before the deadline, but there were also a considerable amount of books returned after the deadline. Utilizing predictive analytics to predict loan lengths would propose TPL information about future incoming book flow, which can be beneficial for multiple operational areas.

Firstly, acquiring information on the daily number of incoming books could enable TPL to enhance capacity planning, an aspect the library has expressed a desire to improve. Capacity planning is generally considered an area that can greatly benefit from the prediction of future outcomes, and forecasting methods are improving. Additionally, the findings of the study revealed a positive correlation between the volume of incoming book flow and the time invested in material handling tasks. Thus, it is reasonable to suggest that obtaining a comprehensive understanding of the incoming book flow would facilitate more effective capacity planning. For TPL, these predictions could materialize to:

- Better planning as the demand for resources becomes more transparent.
- Avoiding burnouts.
- Improving decision-making related to the prioritization of tasks.

Secondly, the prediction of incoming book flow could enhance dynamic sorting in the sorting machine. In the discussion outlined in Section 6.2, it became apparent that the implementation of various modifications to the sorting machine could yield reductions in the time required for material handling tasks, potentially alleviating the mental and physical strain experienced by employees. Consequently, to optimize the performance of the current system, it would be advantageous to introduce multiple alterations. However, due to limitations inherent in the existing machine's capacity, allocating more than 17 bins to accommodate children's books and books destined for other branches while simultaneously maintaining detailed sorting in the remaining bins could pose challenges. To address this issue, a dynamic sorting approach was explored as a viable alternative. A dynamic sorting approach driven by predictive analytics could allow TPL to effectively adapt the sorting machine to the sorting demand, thereby minimizing wasted space within the machine. This could ensure that no bins are dedicated to storing a negligible number of books throughout a given day or an extended period, thus theoretically optimizing the allocation of resources.

Figure 64 summarizes how a library could enhance capacity planning by utilizing predictive analytics:

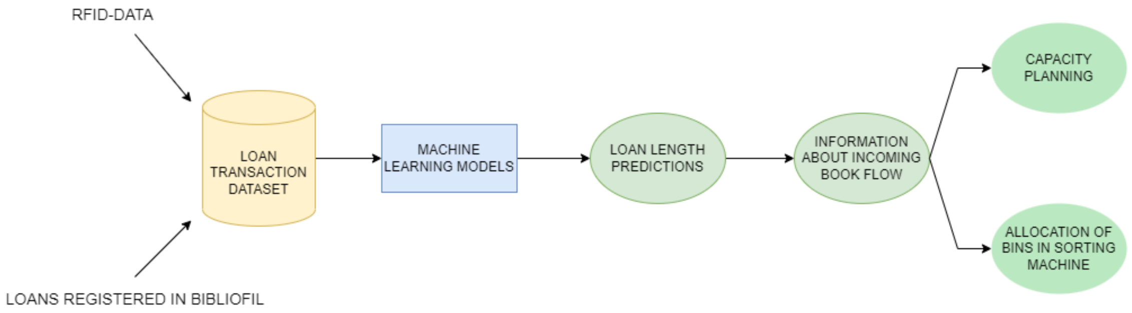

RFID-DATA
LOANS REGISTERED IN BIBLIOFIL
LOAN TRANSACTION DATASET
MACHINE LEARNING MODELS
LOAN LENGTH PREDICTIONS
INFORMATION ABOUT INCOMING BOOK FLOW
CAPACITY PLANNING
ALLOCATION OF BINS IN SORTING MACHINE

Figure 64: How loan data can be utilized to predict incoming book flow and potential application areas.

Nevertheless, even though predictive analytics, in theory, is suitable for these application areas, it is imperative to consider the extent to which TPL truly benefits from the utilization of predictive analytics to enhance its capacity planning and dynamic sorting.

### Predictive Analytics to Enhance Capacity Planning at TPL

Results from the case study indicate that predictive analytics is not a necessary component for achieving a satisfactory level of capacity planning at TPL. These results are substantiated by the staffing plan outlined in the provided case study, which illustrates that TPL maintains a larger workforce from Monday to Thursday in comparison to the remaining days. Taking into account the findings concerning the daily allocation of time towards material handling, this staffing distribution already aligns favorably, as the majority of the time spent on such tasks is concentrated on these weekdays. While these results are based on a limited sample size, the observation is further supported by what librarians have expressed to have experienced earlier with regard to time spent on material handling on different days. Additionally, it is important to consider that the fluctuations in total time spent on material handling in a day are relatively small in relation to the total available manhours. In the sample period examined in this thesis, on the day with the most time spent on material handling relative to manhours available, 20% of manhours were spent on material handling. In comparison, 13% of manhours were spent on material handling on the day with the lowest time spent on this relative to manhours available. The average was just shy of 17%. Consequently, the current staffing plan can be considered satisfactory, and the complexity associated with capacity planning at TPL can be regarded as relatively modest. However, these results might not be indicative of the operations at other libraries, where it occurs larger deviations in the incoming book flow.

Furthermore, to obtain a holistic assessment basis when capacity planning, other factors should be considered in combination with incoming book flow, given that a librarian's workday usually is constituted by material handling, servicing patrons, or planning and arranging events. Information about the events is available without predictive analytics. Time spent on servicing patrons is likely dependent on the number of visitors, which is difficult to predict given available data, and would probably be more accurate by examining historical data. As a result, it is difficult to conclude with predictive analytics being a necessity for capacity planning, at least in the case of TPL.

### Predictive Analytics to Enhance Dynamic Sorting at TPL

To further investigate the advantages of utilizing predictive analytics to alter dynamic sorting, data describing the incoming book flow of three days were analyzed. These days were selected deliberately due to them showing the largest fluctuations in relation to both the volume of incoming books, but also which types of books constituted the book flow. Results indicated that different setups of the sorting machine each day were beneficial. However, the variations between the setups and the corresponding time spent on material handling were relatively minor. In fact, the largest disparity observed in the time spent on material handling across different setups amounted to merely 17 minutes, which accounted for less than a five percent difference in total time spent on material handling. It was found that for all three days, it was beneficial to primarily focus on managing books destined for other branches, reservations, and children's books, as these categories tend to affect time spent on material handling compared to other books, as discussed earlier.

Given the minor discrepancies observed, it can be concluded that the effectiveness of the dynamic sorting approach heavily relies on extremely accurate predictions of the incoming book flow in order to deliver tangible business value. It is plausible that the inflow of books would exhibit greater diversity if seasonal patterns such as Easter and Christmas were taken into consideration, further underscoring the need for dynamic sorting. Unfortunately, only incomplete data from the sorting machine were available for these specific periods.

Consequently, to assess if it was actually feasible to attain predictions of high enough quality to actually propose value with regard to dynamic sorting in TPL, several machine learning models were trained to predict loan lengths, by training on the transaction dataset. Unfortunately, the results obtained from these models were far below acceptable standards. Results actually indicated that the machine learning models barely outperformed constant models. Therefore, it can be considered unlikely that TPL would enhance operations by predicting incoming book flow, which disproves the statement of libraries being able to benefit from predicting incoming book flow, in a practical application.

Another factor that hinders dynamic sorting to benefit from predictive analytics is that very little descriptive information about each book is available in the transaction dataset. This makes it difficult to determine which bin a book belongs to, which is essential information to be able to make informed decisions on the dynamic allocation of bins. Furthermore, the lack of information makes it especially challenging to allocate bins to subgroups of current bins, such as different children's books, and impossible to determine whether a book is reserved. Consequently, the applicability of predictive analytics at TPL is further degraded.

**Summary**

To summarize, utilizing predictive analytics to predict incoming book flow can be considered of little value for TPL. Firstly, it appears like the solution, predictive analytics, is unnecessarily complex in relation to what it yields in terms of improvement in capacity planning and altering the sorting. In addition, predicting incoming book flow proved to be infeasible given the available data. However, libraries that experience more diversity in the input flow of books, both with regard to volume and genres, might be able to benefit from predictive analytics without the predictive models being extremely accurate, which might increase the value of predictive analytics in the discussed application area. However, other measures that could provide libraries with more control of the incoming book flow, which is less complex than predictive analytics, should also be investigated. For instance, every loan could be given a set return date, which the loaner would have to confirm, and be advised to comply with.

## 6.4 Data Quality at TPL

The poor performances obtained by the prediction models correspond to what was described in Nutter (1987), where it was experienced that the libraries' "ability to collect, organize, and manipulate data far outstrips the ability to interpret and apply them". Further, Cheng and Liu (2019) highlights the lack of guidelines and methods for maximizing the utilization of predictive analytics in university libraries, which is likely transferable to public libraries. Consequently, it is interesting to assess the quality of the input data in the prediction models, as it is likely that the poor performances are a product of the collected data not being intended on being utilized as input data for predictive modeling.

From the results presented on data quality, presented in Section 5, it is apparent that the input data actually do hold a high quality with regard to multiple dimensions; it is timely, fairly complete, unique, and holds a satisfactory accuracy. However, for purposes such as predictive analytics, there are several pitfalls with the datasets generated by the library. Firstly, it was experienced that none of the variables in the dataset particularly affect loan length, the target variable in the examined problem. Through correlation analysis, the largest correlating factor was observed to be age, with a correlation coefficient equal to -0.14. With a value this low, the relationship can be said to be negligible. Therefore, it can be stated that the data is currently unfit for the purpose of predicting loan lengths. However, TPL is in possession of more information in their LMS that could increase the performance of predictive models, such as user information. Unfortunately, this information is stored in different datasets described in different formats, which makes for little transparency in the data.

Moreover, even though data seems to be of satisfactory quality with regard to several dimensions, there are still challenges present related to using it for simpler purposes, such as deriving general statistics. For instance, there are several more return transactions, compared to loan transactions, leading to an inconsistency in the data. This phenomenon can be explained by how TPL collects and assigns labels to transactions; some operations that cannot be considered a return, are still assigned the return label. For instance, when a book is scanned and prepared for a reservation, a return transaction will be generated. As a result, statistics about returns derived from the current system should be carefully reviewed, as they are based on data that can be difficult to interpret, thus affecting the reliability of the results.

There are several measures TPL can implement to increase the potential value of their data, with regard to loan- and return transactions. Variables that probably would enhance prediction models' accuracy if they were to be introduced, are presented in the paragraphs below.

**Adding Status in Loan Transactions:** Currently, loan and return transactions are having to be concatenated in order to be able to predict loan lengths. As previously explained, this is a cumbersome process given the discrepancy in loan- and return transactions. Additionally, since the transactions do not possess an ID, wrong transactions might be concatenated, influencing the reliability of the data. To eliminate the need for this concatenation, a status could be introduced in every loan transaction, in addition to the timestamp and location of the return. As a result, transactions solely describing loans and returns would be redundant.

**Deadline of the Loan and Loan Extension:** TPL, and likely other libraries, are operating with loans of different loan lengths. Examples include short loans, loans with standard loan length, one-week loans granted to books with long waiting lists, and lastly longer loans, which are common during the summer months. Therefore, different types of loans are related to different deadlines. Following the logic of Parkinson's Law, the patrons will tend to return books close to the deadline. Even though this law is not scientifically proven in the context of returning borrowed books, it is a generally accepted principle. Furthermore, in Maule et al. (2000), it is investigated how deadlines affect the decision-making of people. The study concludes that people are affected by having a deadline to comply with. Based on these results, it is fair to assume that patrons are affected by the deadline of a loan, which will affect when a book is returned. By analyzing the loan statistics, it is apparent that some loan lengths appear more frequently than others, with the highest peaks appearing at two and four weeks. This also corresponds to deadlines for the different types of loans, supporting Parkinson's Law.

Thus, assigning the loan transactions with their respective deadlines could improve the model predictions. However, the possibility of a loan being extended is also present, and should therefore be considered by the model. For a model to handle these cases, it would be beneficial to introduce a categorical variable that is updated when a loan is extended, in combination with updating the value describing the deadline.

**Reservation Status:** Currently, a reservation is registered as a separate transaction from loans and returns. Consequently, it is not possible to determine if a book is reserved with the current transaction dataset, and thus not possible to determine which bin the book is going to be placed in, in the sorting machine. As a result, it is difficult to utilize loan length predictions to optimize the sorting in the sorting machine.

A feature describing whether the incoming book is reserved or not, in combination with the loaner department of the patron requesting the reservation, would add transparency to which bin the incoming books are going to be placed in. However, this assumes that the reserved books are always collected at the patron's home branch. Furthermore, it also necessitates a change in the policy of the reservation queue. Currently, all reservations are assigned to copy number 0 of the title and the reservation queue is based on First-In-First-Out. For the proposed features to be of value, the patron would have to be assigned to a specific copy of the book, when requesting a reservation. This could, for instance, be achieved by assigning the copy with the nearest upcoming delivery deadline to the patron who requested the reservation first.

In combination with the feature containing the owner department, these proposed features would provide full transparency to which branch books entering the sorting machine, are to be sent. Consequently, it is possible to adapt the sorting based on which branches the incoming books are to be sent to, potentially avoiding allocating ten bins to the branches every day, if not necessary.

**Loaner ID:** Within the existing transaction data, the available information describing borrowers is limited to age, gender, and postal code, thereby providing a scarce understanding of the loaners themselves. Notably, in the study conducted by T. H. Lee and J. W. Lee (2021), results indicated that individual habits affected loan durations within an academic library setting. Hence, the inclusion of more comprehensive personal information, such as a loaner ID for each patron, would prove advantageous in predicting future outcomes at TPL. Loaner ID would provide information about loan history, and thus personal habits. It is widely acknowledged that processes executed by individuals are influenced by their personal characteristics and habits. Consequently, the timing of book returns by patrons is unlikely to deviate from this principle.

Another influential factor affecting loan durations is the practice of bulk loans, which can be conducted either by institutions such as schools or by individual patrons. Upon analyzing the loan and return transaction dataset, a recognizable pattern has emerged wherein loan transactions with equal timestamps and equal personal descriptives, tend to be returned in bulk. However, due to the absence of loaner identification and given that the current data only considers age, gender, and postal code, it is not possible to definitively ascertain whether these occurrences are indeed bulk loans or returns, despite their likelihood. Consequently, to validate these observations, the implementation of loaner IDs would be advantageous. Introducing an attribute that signifies whether a loan is part of a bulk transaction would provide additional information for predictive modeling and thereby contribute to enhancing prediction performance. Lastly, loaner IDs will also give insight into the active loans of a patron; if the loaner ID is present in several loan transactions without appearing in return transactions, the patron has multiple active loans. By possessing this information, TPL could examine if and how this affects loan lengths, and it is possibly something a predictive model would benefit from.

However, for a loaner ID to provide benefits to a predictive model, large amounts of data are likely necessary, as many patrons might not loan numerous books a year. In order for a predictive model to identify patterns related to loaner ID, it would need to see the same IDs multiple times. Furthermore, there are also implications related to implementing an anonymous loaner ID in the datasets. The Norwegian Data Protection Authority has previously blocked this request, indicating that it breaches privacy concerns, even though this is difficult to comprehend. Therefore, this suggestion might be more relevant for libraries outside of Norway.

**More Detailed Information About the Book:** As previously discussed, the execution of a task or process often relies on the characteristics and behaviors of the individual undertaking it. In the context of book loans, the task performed by individuals can be understood as the act of reading and subsequently returning the borrowed book, which is likely also influenced by the specific attributes of the book itself. Therefore, it is logical to include information about the borrowed book in the dataset when attempting to predict loan durations. Presently, the available data concerning the book in loan and return transactions primarily consists of the title ID, copy number, and owner department. However, the title ID is highly specific, making it challenging to identify meaningful patterns. The copy number, although necessary for concatenating loan and return transactions, does not contribute relevant information to a predictive model. The owner department merely denotes whether the book belongs to the children and youth section or not. Additional details regarding the borrowed book can be obtained from the book catalog, which is stored in a separate dataset with a different format. Merging these two datasets requires substantial computational resources and time-consuming processes. Hence, TPL can enhance its data quality by striving to establish greater consistency across its data sources. If the catalog and loan transaction data were in the same format, simple queries could be employed to present loan transactions with additional descriptive information related to the borrowed book. However, the dataset containing information about the book catalog was observed to be scarce, characterized by low uniqueness, and lacked descriptive information about the books which could be valuable for a predictive model, such as the number of pages. Consequently, TPL would likely have to devote notable resources in order to add more descriptive details related to the book in the loan transactions.

**Sorting Groups:** As previously emphasized, it is of great interest to align the sorting of children's books with the temporary shelves in the children and youth area. However, similarly to devoting a bin to every branch every day, it might not be beneficial to devote one bin to every temporary shelf in the children and youth area. However, given the available information in the loan and return transactions, this is difficult to avoid, as it is challenging to map a children's book to its corresponding temporary shelf.

A solution to this problem is adding additional information about the children's books in the transaction data, specifically which temporary shelf it belongs to. This could be applicable to other types of books as well, but likely not necessary in the case of TPL.

### Summary

The performance of prediction models is dependent on the input, as commonly expressed with the expression "garbage in, garbage out". Moreover, a prediction model will always be upper bound by the quality of its input data, which emphasizes the importance of achieving the highest possible quality of input data. In Section 2, the significance of data quality was explained, accompanied by an analogy drawn between data manufacturing and the manufacturing of tangible products. While libraries do not engage in physical product manufacturing, the reverse logistics observed within a library bear resemblance to a production process. In this analogy, incoming books can be compared to raw materials, the material handling tasks involved in ensuring the availability of books to patrons can be viewed as the processing stage, and the end result is the availability of materials for patrons. This process cycle represents an ongoing improvement endeavor for libraries. However, in the case of TPL, the manufacturing aspect of data appears to have been overlooked. The results describing data quality at TPL presented in this thesis have contributed to mapping and highlighting the challenges present in generating and storing data in public libraries. Currently, public libraries, represented by TPL, do not seem prepared to fully take advantage of their amassed data, which decreases the applicability of predictive analytics.

The potential of implementing several features to the transaction dataset has been discussed. Based on these features, it is also possible to engineer new features, as briefly explained when each feature was presented. The table below summarizes the proposed features, independent variables, and their potential value:

| Independent Variable | Potential Value |
|---|---|
| Loan Status | Eliminates the need for concatenating loan- and return transactions. Is dependent on fields describing timestamp of return. |
| Deadline | Provides the model with the deadline of the loan. This has been seen to affect loan length. |
| Loan Extension | Will notify when a deadline should be changed. Also useful to uncover how extensions affect loan length. |
| Reservation Status | Shows if a book is reserved or not. |
| Loaner Department of Requested Reservation | Determines where a reserved book is going to be picked up. |
| Loaner ID | Presents the possibility of investigating how personal habits affect loan length. |
| No. of Pages | Logical that longer books might be loaned for longer periods of time. |
| Genre | Uncovering patterns apparent amongst loans of different book genres. |
| Sorting Groups | More detailed information related to the sorting of children's books. |

Table 17: Summary of the suggestions for improving data quality for predicting loan lengths

However, it is important to consider that the addition of new features does not guarantee improved accuracy for prediction models. To verify the effects of implementing the suggested features, they would have to be added to the datasets, and the performance of prediction models training on the updated data would have to be compared to the results obtained with the currently available data.

Moreover, it is also worth reflecting on whether the benefits proposed by predictive analytics in a library are worth the necessary investment. For instance, it is not certain that a prediction model would be able to predict the incoming book flow in a library, regardless of the data quality. As said by the Danish physicist and Nobel prize-winner, Niels Bohr: "It is difficult to make predictions, especially about the future". In the examined prediction problem, the loan lengths are dependent on individuals who might show entirely different habits, increasing the complexity of the prediction problem. Additionally, it is reasonable to assume that when a patron returns a loan, is dependent on factors that are impossible to comprehend through the analysis of library data, such as personal weekly schedules.

# 7 Conclusion

This chapter presents the conclusions of this thesis, and addresses the four research questions, before describing the thesis' contribution to the field of study and suggestions for further research.

This research aimed to investigate how incoming book flow affects material handling at a library, and how material handling can be made more efficient in a library setting. Lastly, the applicability of predictive analytics in relation to these topics was analyzed. This is a relatively untouched field of inquiry, characterized by a scarcity of existing literature. Consequently, results have been attained through a case study, conducted on Trondheim Public Library.

## 7.1 Research Questions

The following section aims to present the answers for each of the research questions, defined in the introduction chapter.

**RQ1: How does incoming book flow affect material handling in a library?**
Interviews with librarians, alongside on-site observations, helped map the material handling tasks necessary to manage incoming book flow. These material handling tasks were divided into four process steps; material return, sorting, holding, and final shelving. Based on quantitative analysis of time spent on material handling on a daily basis, it can be concluded that the volume of incoming book flow shares a strong relationship with the time spent on material handling tasks in a library, supported by a correlation of 0.762. It was also emphasized that it is not only the volume of the incoming book flow that affects time spent on material handling, but also what constitutes the book flow. However, the material handling related to sorting was found to be unaffected, as a result of the entire sorting process being automated.

**RQ2: How can a library enable more efficient material handling?**
In order to enhance the quality of library services, it is advantageous to minimize the amount of time librarians spend on material handling activities, thereby allowing them to allocate more time to servicing patrons and organizing events. Automation was highlighted as an effective tool to achieve this, but also expensive and resource-consuming to integrate. Therefore, the research focused on specifically examining the sorting process as it was found to have a direct impact on subsequent material handling tasks, and thereby efficiency. The findings revealed that the current sorting approach for children's books and books designated for other branches contributed to an excessive amount of material handling.

Consequently, the study investigated alternative sorting methods that could eliminate the identified excessive workload. However, the key insight derived from this experiment extends beyond the specific alterations implemented in this particular case. It underscored the significance of eliminating redundant manual material handling tasks, resulting in improved efficiency and enhanced employee ergonomics.

**RQ3: How can predictive analytics be applied in a library setting to enhance the planning of material handling?**
This paper highlighted the potential benefits yielded by predictive analytics within capacity planning and bin allocation in the sorting machine, specifically by predicting the incoming book flow. However, in the case of Trondheim Public Library, accurate insights into the book flow were found not to improve capacity planning significantly, likely because of the relatively small deviations in the weekly patterns observed in the incoming book flow, relative to the available manhours. Therefore, the staffing plan currently employed by Trondheim Public Library was found to be of satisfactory quality.

From the analysis of adjusting the sorting in the sorting machine, it became apparent that several alterations to the current system would be beneficial. However, to best utilize the capacity of the machine, it was deemed interesting to examine the potential of a dynamic sorting approach. To investigate the advantages of predictive analytics in dynamic sorting, the study examined the incoming book flow on three selected days with significant deviations. The findings indicated that a dynamic approach would be beneficial, with different optimal setups identified for each day. However, the variations in the optimal setups were minor, likely due to the similarity in the book flow patterns except for volume differences. The study concluded that while dynamic sorting offered benefits, they were not as prominent as they could be in libraries with higher variability in book flow. In this particular case, achieving noteworthy benefits from dynamic sorting would require extremely accurate predictions.

**RQ4: How can a library enhance data quality to make predictive analytics more applicable?**
Further, to investigate how predicting the incoming book flow could be achieved in practice, machine learning models were applied to predict loan lengths. Results suggested that machine learning models are not able to make accurate predictions on the transactional data generated by Trondheim Public Library. Analysis of the dataset indicated that the lack of features affecting the target value, loan length, led to imprecise predictions, and the current data generated can be stated to be unfit for prediction purposes. There are several adjustments that could be implemented in the data-gathering process of a library, that likely would improve data quality. The transactional data would be of more value if further information about the book and patrons were available. Furthermore, more consistent use of the transaction types would be beneficial for preprocessing and data analysis.

## 7.2 Key Takeaways and Contributions

This research has shown that the incoming book flow strongly affects the material handling tasks at a library. Further, the thesis has illustrated how sorting affects subsequent material handling tasks, and revealed how alterations in the sorting can increase efficiency in a material handling system in a library setting. It was found beneficial to alter the sorting in a fashion that enabled the elimination of upcoming material handling tasks. Further, applying predictive analytics for predicting the incoming book flow was found to be difficult, as it required extremely high accuracy to achieve very minor benefits. Therefore, it can be concluded with predictive analytics not being suitable for Trondheim Public Library, with regard to improving capacity planning or optimizing sorting based on incoming book flow. Predictive analytics is an unnecessarily complex solution compared to the potential yield.

This research has contributed to providing detailed insight into how incoming book flow initializes material handling tasks in a library. The contribution extends to knowledge of how the volume of incoming flow determines the workload of material handling in a library. Furthermore, the research has highlighted the sorting process as an area greatly affecting the efficiency of a material handling system in a library setting and proposed alternatives enabling optimized sorting. The findings should be somewhat generalizable to most libraries having a book-sorting machine incorporated into their material handling system. Moreover, this research contributes with general suggestions to how a library can improve its data collection and data quality to better facilitate predictive analytics. These findings are generalizable to libraries currently amassing data mostly for statistical purposes.

## 7.3 Limitations and Further Research

Given the lack of research describing material handling and data analytics in libraries, the findings have mainly been derived based on a single company case study. Consequently, the result can have a tendency to be case-dependent, and it is difficult to determine if the case company, Trondheim Public Library, is representative of other libraries. Furthermore, the research is limited by the amount of quantitative data available. As a result, outliers and anomalies might have had a larger impact on the findings than what is actually the case.

The limitations of this research should be used to define future research in related areas. To further verify or challenge the results attained in this research, it would be necessary to explore other cases. Furthermore, all proposed adjustments to the sorting presented in this thesis are based on sorting continuing to be performed based on genre. For further work, it would be interesting to explore how more drastic changes in sorting would affect the efficiency of material handling. For instance, it would be interesting to review how altering the sorting policy to be based on the books' designated positions in the library, and not genres, would affect efficiency.

Moreover, in order to verify if the proposed feature additions will improve data quality and further investigate if it is possible to improve the predictions of the incoming book flow, it will be necessary to construct test data with additional features. However, continuing research in this field is only suggested if it appears as other libraries might benefit from these predictions more than Trondheim Public Library.

# Bibliography

Abdi, Hervé and Lynne J Williams (2010). 'Principal component analysis'. In: *Wiley interdisciplinary reviews: computational statistics* 2.4, pp. 433–459.

Adèr, Hermanus Johannes (2008). *Advising on research methods: A consultant's companion.* Johannes van Kessel Publishing., pp. 271–304.

Al Shalabi, Luai, Zyad Shaaban and Basel Kasasbeh (2006). 'Data mining: A preprocessing engine'. In: *Journal of Computer Science* 2.9, pp. 735–739.

Ali, Jehad et al. (2012). 'Random forests and decision trees'. In: *International Journal of Computer Science Issues (IJCSI)* 9.5, p. 272.

Anguita, Davide et al. (2012). 'The 'K'in K-fold cross validation'. In: *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN).* i6doc. com publ, pp. 441–446.

Apple, JM (1972). *Material Handling System Design.* New York, Ronald Press Co.

Arlitsch, Kenning and Bruce Newell (2017). 'Thriving in the age of accelerations: a brief look at the societal effects of artificial Intelligence and the opportunities for libraries'. In: *Journal of Library Administration* 57.7, pp. 789–798.

Asuero, Agustin Garcia, Ana Sayago and AG González (2006). 'The correlation coefficient: An overview'. In: *Critical reviews in analytical chemistry* 36.1, pp. 41–59.

Ayre, Lori Bowen (2012). 'Library RFID systems for identification, security, and materials handling'. In: *Library Technology Reports* 48.5, pp. 9–16.

Baba, Kensuke, Toshiro Minami and Tetsuya Nakatoh (Dec. 2016). 'Predicting Book Use in University Libraries by Synchronous Obsolescence'. In: *Procedia Computer Science* 96, pp. 395–402. DOI: 10.1016/j.procs.2016.08.082.

Bakkali, Hajira, Abdellah Azmani and Abdelhadi Fennan (Dec. 2013). 'Dynamic Allocation of Products to Storage Areas in the Warehouse'. In: *International Journal of Computer Applications* 84, pp. 36–43. DOI: 10.5120/14663-2974.

Barriball, K Louise and Alison While (1994). 'Collecting data using a semi-structured interview: a discussion paper'. In: *Journal of Advanced Nursing-Institutional Subscription* 19.2, pp. 328–335.

Berg, Jeroen (Aug. 1999). 'A literature survey on planning and control of warehousing systems'. In: *IIE Transactions* 31, pp. 751–762. DOI: 10.1023/A:1007606228790.

Botchkarev, Alexei (2018). 'Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology'. In: *arXiv preprint arXiv:1809.03006.*

— (2019). 'A new typology design of performance metrics to measure errors in machine learning regression algorithms'. In: *Interdisciplinary Journal of Information, Knowledge, and Management* 14, p. 45.

Brodley, Carla E, Mark A Friedl et al. (1996). 'Identifying and eliminating mislabeled training instances'. In: *Proceedings of the National Conference on Artificial Intelligence*, pp. 799–805.

Cai, Jie et al. (2018). 'Feature selection in machine learning: A new perspective'. In: *Neurocomputing* 300, pp. 70–79.

Cameron, A Colin and Frank AG Windmeijer (1996). 'R-squared measures for count data regression models with applications to health-care utilization'. In: *Journal of Business & Economic Statistics* 14.2, pp. 209–220.

— (1997). 'An R-squared measure of goodness of fit for some common nonlinear regression models'. In: *Journal of econometrics* 77.2, pp. 329–342.

Cebeci, Zeynel and Figen Yildiz (2017). 'Comparison of Chi-square based algorithms for discretization of continuous chicken egg quality traits'. In: *Journal of Agricultural Informatics* 8.1.

Chai, Tianfeng and Roland R Draxler (2014). 'Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature'. In: *Geoscientific model development* 7.3, pp. 1247–1250.

Charbuty, Bahzad and Adnan Abdulazeez (2021). 'Classification based on decision tree algorithm for machine learning'. In: *Journal of Applied Science and Technology Trends* 2.01, pp. 20–28.

Cheng, Yanping and Qingyu Liu (2019). 'Process and application of data mining in the university library'. In: *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*. IEEE, pp. 123–127. ISBN: 1728112826.

Chu, Xu et al. (2016). 'Data cleaning: Overview and emerging challenges'. In: *Proceedings of the 2016 international conference on management of data*, pp. 2201–2206.

Clark, Peter and Tim Niblett (1989). 'The CN2 induction algorithm'. In: *Machine learning* 3.4, pp. 261–283.

Coyle, John Joseph, Edward J Bardi and C John Langley (1992). *The management of business logistics: A supply chain perspective.* South-Western/Thomson Learning.

Cunningham, Pádraig (2008). 'Dimension reduction'. In: *Machine learning techniques for multimedia: Case studies on organization and retrieval*, pp. 91–112.

De Myttenaere, Arnaud et al. (2016). 'Mean absolute percentage error for regression models'. In: *Neurocomputing* 192, pp. 38–48.

Deo, Ms Gouri S (2020). 'IoT-enabled Library Management System with Predictive Analysis of Resource Usage Data using Machine Learning for the Qualitative Up-gradation'. In: *Diss. Amity University.*

Developers, scikit-learn (2023). *sklearn.svm.SVR.* Accessed: 2023-06-01.

Dewey, Melvil (1876). *A classification and subject index, for cataloguing and arranging the books and pamphlets of a library.* Brick row book shop, Incorporated.

Dietterich, Tom (1995). 'Overfitting and undercomputing in machine learning'. In: *ACM computing surveys (CSUR)* 27.3, pp. 326–327.

Dougherty, James, Ron Kohavi and Mehran Sahami (1995). 'Supervised and unsupervised discretization of continuous features'. In: *Machine learning proceedings 1995*. Elsevier, pp. 194–202.

Eckerson, Wayne W (2007). 'Predictive analytics'. In: *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report* 1, pp. 1–36.

Eisenhardt, Kathleen M (1989). 'Building theories from case study research'. In: *Academy of management review* 14.4, pp. 532–550.

Emmanuel, Tlamelo et al. (2021). 'A survey on missing data in machine learning'. In: *Journal of Big Data* 8.1, pp. 1–37.

Feng, De-Cheng et al. (2020). 'Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach'. In: *Construction and Building Materials* 230, p. 117000.

Ferguson, Stuart J and Rodney Hebels (2003). 'Library management systems'. In: *Computers for Librarians (Third Edition)*. Chandos Publishing, pp. 111–142.

Fernandez, Jeffrey E (1995). 'Ergonomics in the workplace'. In: *Facilities* 13.4, pp. 20–27.

Fernstad, Sara Johansson (2019). 'To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization'. In: *Information Visualization* 18.2, pp. 230–250.

Garcia, Salvador et al. (2016). 'Big data preprocessing: methods and prospects'. In: *Big Data Analytics* 1.1, pp. 1–22.

Gudivada, Venkat, Amy Apon and Junhua Ding (2017). 'Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations'. In: *International Journal on Advances in Software* 10.1, pp. 1–20.

Guyon, Isabelle and André Elisseeff (2003). 'An introduction to variable and feature selection'. In: *Journal of machine learning research* 3.Mar, pp. 1157–1182.

Hastie, Trevor et al. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hazen, Benjamin T et al. (2014). 'Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications'. In: *International Journal of Production Economics* 154, pp. 72–80.

Heaton, Jeff (2016). 'An empirical analysis of feature engineering for predictive modeling'. In: *SoutheastCon 2016*. IEEE, pp. 1–6.

Heinrich, Bernd et al. (2018). 'Requirements for data quality metrics'. In: *Journal of Data and Information Quality (JDIQ)* 9.2, pp. 1–32.

Hicks, Steven A et al. (2022). 'On evaluation metrics for medical applications of artificial intelligence'. In: *Scientific Reports* 12.1, p. 5979.

Hiroyuki, Hirano (2012). *JIT Business Research*. URL: https://www.jit-ken.co.jp/ (visited on 8th Dec. 2022).

Huang, Jui-Chan et al. (2020). 'Application and comparison of several machine learning algorithms and their integration models in regression problems'. In: *Neural Computing and Applications* 32.10, pp. 5461–5469.

Hurmerinta-Peltomäki, Leila and Niina Nummela (2006). 'Mixed methods in international business research: A value-added perspective'. In: *Management International Review* 46, pp. 439–459.

Hye, AK Mahbubul, Engku M Nazri and Nurakmal Ahmad Mustaffa (2019). 'The library supply chain model: A brief'. In: *Int. J Sup. Chain. Mgt Vol* 8.6, p. 32.

IBM (2020). *Supervised learning.* URL: https://www.ibm.com/cloud/learn/supervised-learning (visited on 7th Nov. 2022).

— (2023). *What is data quality?* URL: https://www.ibm.com/topics/data-quality (visited on 20th May 2023).

IEA (2000). *What Is Ergonimics(HFE)?* Accessed: 2022-12-19.

Ij, H (2018). 'Statistics versus machine learning'. In: *Nat Methods* 15.4, p. 233.

Institute, Material Handling (2023). *Material Handling.* Accessed: 2023-05-06.

Iqbal, Naeem et al. (2020). 'Toward Effective Planning and Management Using Predictive Analytics Based on Rental Book Data of Academic Libraries'. In: *IEEE Access* 8, pp. 81978–81996. DOI: 10.1109/ACCESS.2020.2990765.

John, George H (1995). 'Robust Decision Trees: Removing Outliers from Databases.' In: *KDD*. Vol. 95, pp. 174–179.

Jones-Farmer, L Allison, Jeremy D Ezell and Benjamin T Hazen (2014). 'Applying control chart methods to enhance data quality'. In: *Technometrics* 56.1, pp. 29–41.

Jordan, Michael I and Tom M Mitchell (2015). 'Machine learning: Trends, perspectives, and prospects'. In: *Science* 349.6245, pp. 255–260.

Karlsson, Christer (2010). *Researching operations management.* Routledge.

Kay, Michael G (2012). 'Material handling equipment'. In: *Fitts Dept. of Industrial and Systems Engineering North Carolina State University* 65.

Klein, Katherine J. and Joann Speer Sorra (1996). 'The Challenge of Innovation Implementation'. In: *The Academy of Management Review* 21.4, pp. 1055–1080. ISSN: 03637425. URL: http://www.jstor.org/stable/259164 (visited on 12th Nov. 2022).

Klimushkin, Mikhail, Sergei Obiedkov and Camille Roth (2010). 'Approaches to the selection of relevant concepts in the case of noisy data'. In: *Formal Concept Analysis: 8th International Conference, ICFCA 2010, Agadir, Morocco, March 15-18, 2010. Proceedings 8.* Springer, pp. 255–266.

Kothari, Chakravanti Rajagopalachari (2004). *Research methodology: Methods and techniques.* New Age International.

Kumar, Sunil and Ilyoung Chong (2018). 'Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states'. In: *International journal of environmental research and public health* 15.12, p. 2907.

Kusiak, Andrew (2001). 'Feature transformation methods in data mining'. In: *IEEE Transactions on Electronics packaging manufacturing* 24.3, pp. 214–221.

Labajo, Ellen M (2017). 'Occupational ergonomics in the library workplace'. In: *University of the Visayas-Journal of Research* 11.1, pp. 53–60.

Learning, Google Developers Machine (2022a). *Bucketing.* Accessed: 2023-05-03.

— (2022b). *Introduction to Transforming Data.* Accessed: 2023-05-03.

— (2022c). *Normalization.* Accessed: 2023-05-03.

Lee, Carman KM et al. (2018). 'Design and application of Internet of things-based warehouse management system for smart logistics'. In: *International Journal of Production Research* 56.8, pp. 2753–2768.

Lee, Ga Young et al. (2021). 'A survey on data cleaning methods for improved machine learning model performance'. In: *arXiv preprint arXiv:2109.07127.*

Lee, Tae Ho and Jae Woo Lee (2021). 'Self-organized human behavioral patterns in book loans from a library'. In: *Physica A: Statistical Mechanics and its Applications* 563, p. 125473.

Lian, Defu et al. (2016). 'Mutual Reinforcement of Academic Performance Prediction and Library Book Recommendation'. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1023–1028. DOI: 10.1109/ICDM.2016.0130.

Livingston, Frederick (2005). 'Implementation of Breiman's random forest machine learning algorithm'. In: *ECE591Q Machine Learning Journal Paper*, pp. 1–13.

Ludbrook, John (2008). 'Outlying observations and missing values: how should they be handled?' In: *Clinical and experimental pharmacology & physiology* 35.5-6, pp. 670–678.

Makridakis, Spyros, Evangelos Spiliotis and Vassilios Assimakopoulos (2018). 'Statistical and Machine Learning forecasting methods: Concerns and ways forward'. In: *PloS one* 13.3, e0194889.

Maule, A.John, G.Robert J Hockey and L Bdzola (2000). 'Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy'. In: *Acta Psychologica* 104.3, pp. 283–301. ISSN: 0001-6918. DOI: https://doi.org/10.1016/S0001-6918(00)00033-0.

McCullagh, Peter (2002). 'What is a statistical model?' In: *The Annals of Statistics* 30.5, pp. 1225–1310.

McLeod, Saul (2019). 'Qualitative vs Quantitative Research: Methods & Data Analysis'. In.

Meyer, David and FT Wien (2015). 'Support vector machines'. In: *The Interface to libsvm in package e1071* 28, p. 20.

Mitchell, Tom M and Tom M Mitchell (1997). *Machine learning.* Vol. 1. 9. McGraw-hill New York.

Mushtaq, Sana (2019). *Data preprocessing in detail.* URL: https://developer.ibm.com/articles/data-preprocessing-in-detail/ (visited on 29th Nov. 2022).

Natekin, Alexey and Alois Knoll (2013). 'Gradient boosting machines, a tutorial'. In: *Frontiers in neurorobotics* 7, p. 21.

Noble, William S (2006). ‘What is a support vector machine?’ In: *Nature biotechnology* 24.12, pp. 1565–1567.

Nutter, Susan K (1987). ‘Online systems and the management of collections: Use and implications’. In: *Advances in Library Automation Networking* 1, pp. 125–149.

Obaid, Hadeel S, Saad Ahmed Dheyab and Sana Sabah Sabry (2019). ‘The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning’. In: *2019 9th annual information technology, electromechanical engineering and microelectronics conference (iemecon)*. IEEE, pp. 279–283.

Oodi (2019). *Oodi robots named after children's book characters.* Accessed: 2023-05-22.

Pedregosa, F. et al. (2011). ‘Scikit-learn: Machine Learning in Python’. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peterson, Eric D (2019). ‘Machine learning, predictive analytics, and clinical practice: can the past inform the present?’ In: *Jama* 322.23, pp. 2283–2284.

Pipino, Leo L, Yang W Lee and Richard Y Wang (2002). ‘Data quality assessment’. In: *Communications of the ACM* 45.4, pp. 211–218.

Pop, Corina and Gabriela Mailat (2011). ‘Automated Material Handling Systems (AMHS) in libraries and archives Automated Storage/retrieval and Return/sorting Systems’. In: *Proceedings of the 12th WSEAS international conference on Neural networks, fuzzy systems, evolutionary computing & automation.* Citeseer, pp. 189–194.

Queirós, André, Daniel Faria and Fernando Almeida (2017). ‘Strengths and limitations of qualitative and quantitative research methods’. In: *European journal of education studies.*

Raschka, Sebastian (2018). ‘Model evaluation, model selection, and algorithm selection in machine learning’. In: *arXiv preprint arXiv:1811.12808.*

Ribeiro, Rita P and Nuno Moniz (2020). ‘Imbalanced regression and extreme value prediction’. In: *Machine Learning* 109, pp. 1803–1835.

Rojas, Raúl et al. (2009). ‘AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting’. In: *Freie University, Berlin, Tech. Rep.*

Roodbergen, KJ and R Meller (2004). ‘Storage assignment policies for warehouses with multiple cross aisles’. In: *Progress in Material Handling Research* 431, p. 441.

Seyedan, Mahya and Fereshteh Mafakheri (2020). ‘Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities’. In: *Journal of Big Data* 7.1, pp. 1–22.

Sgarbossa, Fabio et al. (2020). ‘Human factors in production and logistics systems of the future’. In: *Annual Reviews in Control* 49, pp. 295–305.

Silver, Edward Allen, David F Pyke, Rein Peterson et al. (1998). *Inventory management and production planning and scheduling.* Vol. 3. Wiley New York.

Silverstein, Craig and Stuart M Shieber (1996). ‘Predicting individual book use for off-site storage using decision trees’. In: *The Library Quarterly* 66.3, pp. 266–293.

Sindhu, V, S Nivedha and M Prakash (2020). ‘An empirical science research on bioinformatics in machine learning’. In: *Journal of Mechanics Of Continua And Mathematical Sciences.*

Sitanggang, Imas Sukaesih et al. (2010). ‘Sequential pattern mining on library transaction data’. In: *2010 International Symposium on Information Technology.* Vol. 1, pp. 1–4. DOI: 10.1109/ITSIM.2010.5561316.

Sobhani, Ahmad, MIM Wahab and W Patrick Neumann (2017). ‘Incorporating human factors-related performance variation in optimizing a serial system’. In: *European Journal of Operational Research* 257.1, pp. 69–83.

Tayi, Giri Kumar and Donald P Ballou (1998). ‘Examining data quality’. In: *Communications of the ACM* 41.2, pp. 54–57.

Teng, Choh-Man (1999). ‘Correcting Noisy Data.’ In: *ICML*. Vol. 99. Citeseer, pp. 239–248.

Thibodeau, Patricia L and Steven J Melamut (1995). ‘Ergonomics in the electronic library.’ In: *Bulletin of the Medical Library Association* 83.3, p. 322.

Tsuji, Keita et al. (2014). ‘Book recommendation based on library loan records and bibliographic information’. In: *Procedia-social and behavioral sciences* 147, pp. 478–486.

Uppal, Veepu and Gunjan Chandwani (July 2013). ‘An Empirical Study of Application of Data Mining Techniques in Library System’. In: *International Journal of Computer Applications* 74, pp. 42–46. DOI: 10.5120/12933-0008.

Verleysen, Michel and Damien François (2005). ‘The curse of dimensionality in data mining and time series prediction’. In: *Computational Intelligence and Bioinspired Systems: 8th International Work-Conference on Artificial Neural Networks, IWANN 2005, Vilanova i la Geltrú, Barcelona, Spain, June 8-10, 2005. Proceedings 8.* Springer, pp. 758–770.

Vijayakumar, Vivek et al. (2022). ‘Framework for incorporating human factors into production and logistics systems’. In: *International Journal of Production Research* 60.2, pp. 402–419.

Wang, Richard Y, Veda C Storey and Christopher P Firth (1995). ‘A framework for analysis of data quality research’. In: *IEEE transactions on knowledge and data engineering* 7.4, pp. 623–640.

Wang, Sun-Chong (2003). ‘Artificial neural network’. In: *Interdisciplinary computing in java programming.* Springer, pp. 81–100.

Willmott, Cort J and Kenji Matsuura (2005). ‘Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance’. In: *Climate research* 30.1, pp. 79–82.

Wolpert, David H and William G Macready (1997). ‘No free lunch theorems for optimization’. In: *IEEE transactions on evolutionary computation* 1.1, pp. 67–82.

Yang, Li and Abdallah Shami (2020). ‘On hyperparameter optimization of machine learning algorithms: Theory and practice’. In: *Neurocomputing* 415, pp. 295–316.

Yang, Shih-Ting and Ming-Chien Hung (2012). ‘A model for book inquiry history analysis and book-acquisition recommendation of libraries’. In: *Library Collections, Acquisitions, & Technical Services* 36.3-4, pp. 127–142.

Zheng, Alice and Amanda Casari (2018). *Feature engineering for machine learning: principles and techniques for data scientists.* ” O’Reilly Media, Inc.”

NTNU

Norwegian University of
Science and Technology