Norwegian University
of Life Sciences

Master's Thesis 2024 30 ECTS
Faculty of Science and Technology

Prediction of destination choice in transport modelling with mobile phone data and machine learning models: Experiences from Tønsberg

Kristian Olsen
Industrial Economics

Abstract

The development of transportation infrastructure is a prolonged and complex process that requires in depth planning. The planning process is costly, and further investments for construction of infrastructure is even larger. Transport modelling predicts future traffic flow to appropriately dimension the required infrastructure. The study addresses the prediction of destination choice in the four-step model (4SM) in transport modelling with combining mobile phone data and machine learning algorithms.

Traditionally, destination choice models relied on theory-based discrete choice models using data from travel surveys. However, emerging of Big Data and advances in artificial intelligence have facilitated the use of more representative data, possible for enhancing predictive accuracy and potentially reducing costs and risks associated with the over- or under-construction of infrastructure. Previous research has focused on either the application of mobile phone data or artificial intelligence independently in transport modelling. This research aims to investigate the performance by combining these two novel approaches in transport modelling.

The selection of algorithms tested includes Logistic Regression, Support Vector Machine, Random Forest and Naïve Bayes. They are tested within a basic analytical pipeline to assess performance. Performance evaluations were conducted using a five-fold cross-validation on performance metrics, and through a multi-criteria decision analysis (MCDA) to consider both qualitative and quantitative criteria for model performance. Results indicate robust performance across all algorithms on mobile phone data, with Random Forest performing best, considering both quantitative and qualitative metrics, achieving a prediction accuracy of 0.943. Naïve Bayes and Support Vector Machine followed with 0.925 and 0.895, respectively, while Logistic Regression achieved 0.832. Simpler hyperparameters yielded better results for Support Vector Machine and Logistic Regression, whereas Random Forest utilized more complex hyperparameters for best performance.

These findings suggest that integrating mobile phone data with machine learning algorithms holds substantial promise for enhancing the prediction of destination choices. Future research should explore the model's applicability across different geographic contexts and further steps towards its implementation and deployment.

Keywords: transport modelling, travel demand, mobile phone data, machine learning, artificial intelligence, Big Data, destination choice.

Sammendrag

Utviklingen av transportinfrastruktur er en langvarig og kompleks prosess som krever grundig planlegging. Planleggingen av infrastruktur er svært kostnadskrevende, og ytterligere investeringer kreves til konstruksjon av infrastrukturen. Transportmodellering forutsier fremtidig trafikkflyt for å dimensjonere den nødvendige infrastrukturen på en passende måte. Denne oppgaven tar for seg prediksjon av destinasjonsvalg i firestrinnsmodellen (4SM) i transportmodellering ved å kombinere mobildata og maskinlæringsalgoritmer.

Tradisjonelt har modeller for valg av destinasjon støttet seg på teoribaserte, diskrete valgmodeller som bruker data fra reisevaneundersøkelser. Imidlertid har fremveksten av nye, store datakilder og fremskritt innen kunstig intelligens muliggjort bruken av mer representative data, noe som kan bedre prediksjonsnøyaktiget og potensielt redusere kostnader forbundet med over- eller underdimensjonering av infrastruktur. Tidligere forskning i transportmodellering har satt søkelys på enten bruk av mobildata eller kunstig intelligens uavhengig av hverandre. Denne forskningen sikter mot å undersøke ytelsen ved å kombinere disse to nye tilnærmingene til transport modellering.

Utvalget av algoritmer som testes er *Logistic Regression, Support Vector Machine, Random Forest* og *Gaussian Naïve Bayes*. De er testet innenfor en grunnleggende analytisk modelleringsstruktur for å vurdere den enkeltes prestasjon. Prestasjonsevalueringen ble gjennomført ved hjelp av en kryssvalidering på kvalitative kriterier og gjennom en flermålsanalyse (MCDA) for å vurdere både kvalitative og kvantitative kriterier. Resultatene viser lovende ytelse på tvers av alle algoritmer på mobildata, med Random Forest som den best presterende, også når hensyntatt kvantitative og kvalitative mål. Den oppnår er prediksjonsnøyaktighet på 0.943. Naïve Bayes og Support Vector Machine fulgte etter med henholdsvis 0.925 og 0.895, mens Logistic Regression oppnådde 0.832. Enklere hyperparametre ga bedre resultater for Support Vector Machine og Logistic Regression, mens Random Forest brukte mer komplekse hyperparametre for best ytelse.

Disse funnene antyder at integrering av mobildata med maskinlæringsalgoritmer har betydelig potensial for å forbedre prediksjonen av destinasjonsvalg. Fremtidig forskning bør utforske modellens anvendelighet over forskjellige geografiske områder og utforske ytterligere skritt mot implementering av modellene.

Nøkkelord: Transportmodellering, reiseetterspørsel, mobildata, maskinlæring, kunstig intelligens, destinasjonsvalg

Preface

This master's thesis concludes my five-year Master of Science in Industrial Economics at the Norwegian University of Life Sciences, Faculty of Science and Technology.

The purpose of the thesis has been to research how artificial intelligence models can utilize mobile phone data for destination choice predictions in transport modelling. I would like to express my gratitude to my supervisor, associate professor Øyvind Lervik Nilsen, and co-supervisor, associate professor Tor Kristian Stevik, for valuable guidance and insightful discussions. Your contributions and feedback to the research has been truly appreciated. I am also grateful to the industry partner Rambøll Norway for providing real-world data and industry insight that enriched this study.

Special thanks to the Eik Lab community for their camaraderie whilst studying and working on innovation projects. I equally cherish the memories from Heimen and Laget på Ås, for valuable time spent together.

Lastly, I would like to direct my profound appreciation to my wife, for her immense support in my studies. Her encouragement has been invaluable for both studies and all extracurricular activities I've been involved in throughout my studying years.

Kristian Olsen

Ås, May 2024

Table of Content

List of Figures

List of tables

Terminology

Term	Description
Trip	A trip is a travel between at least two zones.
Trip generation	This refers to the process of estimating the number of trips originating in or destined for a particular area within a given time-period. Trip generation analysis focuses on understanding how many trips will be made, often based on variables such as land use, demographic characteristics, and economic factors. It's the first step in the four-step transportation forecasting process, which also includes trip distribution, mode choice, and route assignment.
Trip distribution	Trip distribution models how trips are dispersed across the transport network. This step involves predicting the destination of trips originating from each zone, based on factors like distance, travel time, and the attractiveness of potential destinations. It provides insight into the flow of trips between different parts of the network.
Mode choice	Mode choice analysis predicts the transportation modes (e.g., car, public transit, walking, cycling) that individuals will choose for their trips, based on factors such as cost, time, convenience, and personal preferences. It is an essential step in transport modelling because it helps planners understand the demand for different modes of transport and plan accordingly.
Trip attraction	Trip attraction is similar to trip generation but focuses on the destinations. It estimates the number of trips that will be attracted to different locations within the area being studied. These locations could be commercial centers, workplaces, schools, or any other places that attract people for various reasons.
Trip production	This term is often used interchangeably with trip generation. It specifically refers to the estimation of the number of trips originating from a particular area. Like trip generation, it focuses on the origins of trips and is influenced by the socio-economic characteristics of the population in the area.
DataFrame	A DataFrame is a two-dimensional, table-like data structure with rows and columns, commonly used in data analysis and machine learning for organizing and manipulating data.
Data quality	Data quality refers to the accuracy, completeness, reliability, and relevance of data within the context of its intended use. High-quality data is essential for making informed decisions, accurate analyses, and developing effective machine learning models.
Overfitting	Overfitting is when the machine learning model performs well on the training data, but poor on the test data. Implying that it has been trained too much and becoming too specific and complex.
Underfitting	Underfitting is when the machine learning model is not trained enough and performs bad on training data. Often a sign on the model being too simple to find the pattern in the data.

Abbreviations

AI: Artificial Intelligence

ASD: Active Signaling Data

CDR: Call Detail Records

CRISP-DM: Cross-Industry Standard Process for Data Mining

DCM: Discrete Choice Model

MCC: Matthew's Correlation Coefficient

ML: Machine Learning

MPD: Mobile Phone Data

RF: Random Forest

NaN-values: Not-a-Number values

PCA: Principal Component Analysis

PSD: Passive Signaling Data

MCDA: Multi-criteria decision analysis

REC: Recall

PRE: Precision

TS: Travel Survey

WBS: Work Breakdown Structure

Chapter 1

Introduction

Transport infrastructure challenges are continuously turning more complex. The projects have increasing demands for better utilization of limited resources, limiting CO_2-emissions and conduct minimal disruption to undisturbed nature. The primary objective of infrastructure development is to fulfill societal transportation needs society needs in a manner that is sustainable, efficient, and safe (Odeck and Welde, 2015). Substantial governmental investments are often associated with these projects, thereby imposing a moral obligation to allocate resources in a manner that maximizes societal benefits.

In the planning phase of new infrastructure projects, transport models are used to simulate the predicted human traffic flows across various zones, a process known as destination choice. This study investigates the possibilities for increased prediction accuracy of travel destination choice by transitioning from traditional logit models to artificial intelligence-based models which leverage mobile phone data. The research is conducted in collaboration with the Smart Mobility department at Rambøll Norway, a consulting engineer firm.

This chapter outlines the background for the thesis, followed by research scope and knowledge gaps, the problem statement and research questions, and concludes with addressing limitations of the research.

1.1 Background

Our society fundamentally relies on infrastructure to facilitate the transportation of goods on services. Enabling the possibility to trade and meet various needs with the distribution of resources across the country. The public road network in Norway is governed by the Norwegian Public Roads Administration (NPRA) and extends over 95 200 km, more than twice the circumference of the Earth times (Hoff and Nordahl, 2024). Showcasing that there is a considerable volume of projects related to infrastructure development.

Transport modelling serves to analyze and predict the movement of people and goods across regions. It's an important tool for transport engineers to understand existing conditions, forecast future travel demand, and assess the impacts of new infrastructure projects. Traditionally, discrete choice models have been used for forecasting destination choice and mode choice in transport modelling. These models estimate the probability of an individual to selecting a specific option, derived from socioeconomic characteristics and the relative attractiveness of the options (Ortuzár and Willumsen, 1994).

The recently published Norwegian national transport plan for 2025-2036 (NTP) by the Norwegian Ministry of Transport (2024) underscores the potential of artificial intelligence in the transport sector. According to Utne et al. (2022) maintaining an efficient and secure road infrastructure is a matter of community safety and national security. This infrastructure is increasingly recognized as not only physical but also digital. The NTP strategy emphasizes the need to take more usage of the large amounts of data accumulated in the sector (Norwegian Ministry of Transport, 2024). To enable reliable access to infrastructure, it is imperative to leverage the digital opportunities for information gathering as a step in the digital transformation. With the availability of mobile phone data and advancements in artificial intelligence, it is possible to further enhance transport modeling in improving the precision and efficacy of travel pattern predictions. Anda, Erath and Fourie (2017) suggest that future research should focus on exploring machine learning techniques to effectively process mobile phone data.

With the emerging of computers and digital computation power it has opened new capabilities within data-driven decision-making and predictive modelling. These models are typically executed using industry-specific transport model software. However, despite improvements in software, the type of data utilized has remained constant. Caceres, Romero and Benitez (2020) along with Svaboe (2024), indicate that the dominant method for collecting data in transport modelling through destination choice relies on travel surveys (TS). These surveys often suffer from low response rates yet are aggregated to represent the general population, meaning that a limited number of responses can significantly influence the development of transport models used

in planning. According to §1a of the Norwegian Road Law (1963) the objective is to ensure that all infrastructure serves the road users' interests. Indicating the necessity for infrastructure projects to be based on current or future needs. Consequently, there is an ongoing search for better representation of the general population, with the utilization of mobile phone data as one of the potential approaches.

Rambøll's support for this study underscores the insufficient exploration of this field, highlighting the significant benefits of evolving towards a more data-driven approach in the transport infrastructure sector. Rahnasto (2022) investigated the application of machine learning in transport modelling, utilizing data from a travel survey to predict destination choices, yielding promising results supporting the use of machine learning. With the newfound accessibility to extensive mobile phone data, further research is exploring whether this can enhance the gains of machine learning, potentially serving as a complement or alternative traditional discrete choice models.

1.2 Research Scope and Knowledge Gaps

Existing literature has demonstrated a growing interest in the integration of mobile phone data and AI models into transport modeling, as noted by Anda, Erath and Fourie (2017). Essadeq and Janik (2021) investigated the standalone validation of mobile phone data, though their study lacked incorporation into a modeling framework and did not utilize machine learning. This was a French case study, and there is a need to better understand how this can be used in a Norwegian context. Additionally, Shoman et al. (2023) identifies significant knowledge gaps in transport modelling, specifically relating to freight transport. They suggest that the availability of Big Data could potentially lead to a full replacement of traditional models, or at least cover up for the weaknesses and strengthen traditional transport models. This thesis aims to explore the application of mobile phone data in Tønsberg for constructing AI-driven models to estimate destination choices among individuals. It will examine a range of fundamentally distinct machine learning algorithm to determine which algorithm offers optimal performance when applied to mobile phone data, and whether these can enhance the precision of destination choice models.

1.3 Problem Statement

Considering advancements in technology and the accessibility to new types of data, a novel opportunity space has emerged within the field of transport modeling. Literature identifies a notable knowledge gap in how to connect both the emerging artificial intelligence-based models and the emerging of Big Data. Among the various sources of Big Data, mobile phone data is distinguished as possessing significant potential.

The emerging of both Artificial intelligence-based models and Big Data, particularly through mobile phone data, have unveiled new opportunities for transport modelling. Traditional data sources, which are typically highly aggregated, costly, and often provide a questionable representation of the population, are increasingly challenged by the need to leverage new technologies. The transport industry faces the to further understand the underlying potential in mobile phone data, and utilize the technology advancements in a cost-efficient manner, aiming to improve the accuracy of predictions, models, and representations of the population.

1.3.1 Research Questions

To gain insight in how artificial intelligence models can be applied in transport modelling and explore the problem statement, the research questions to be investigated for this thesis are:

RQ1: *Are AI-based models capable of utilizing mobile phone data to predict destination choices?*
RQ2: *What opportunities and challenges arise when integrating mobile phone data with artificial intelligence in transport modelling?*

1.4 Research Limitations

Numerous factors that influencing real-world scenarios present challenges in covering them within a single model or research project. Given the research on mobile phone data in transport modelling for destination choice is novel, it is advantageous to focus on a smaller application scope of this data with machine learning techniques. Due to further constraints, also imposed by availability of the data, the following limitations is considered in the thesis:

- The research is concentrated to road transport, excluding rail transport.
- The data originates from Tønsberg municipality in Norway.
- The data is sourced from the Norwegian telecommunications operator Telia.
- The thesis concentrates on trip distribution, detailed further in section 2.2.2.

Given the dataset specific to Tønsberg it may be influenced by local cultural factors, potentially limiting its representativeness of other cities, both within Norway and globally. In Norway the two main telecommunications operators, Telia and Telenor, hold market shares of 30,9% and 54,8%, respectively (Tefficient, 2022). Although this might impact the data observations, it is not deemed critical to the research outcomes. For optimal performance, a machine learning algorithm to be requires a clearly defined objective; set to be prediction of the trip distribution (destination choice) within Tønsberg's population. This focus narrows the applicability of the findings to other aspects of transport modelling for how mobile phone data can be utilized.

1.5 Structure of Thesis

The structure of the thesis adheres to the AIMRAD framework, which includes the components *abstract, introduction, materials and methods, results* and *discussion* (Cargill and O'Connor, 2009). Figure 1 presents the Work Breakdown Structure (WBS), a strategic tool employed to scope the research (Rolstadås *et al.*, 2021). The thesis is organized into six chapters, each containing associated subchapters, made to address the formulated problem statement and research questions.

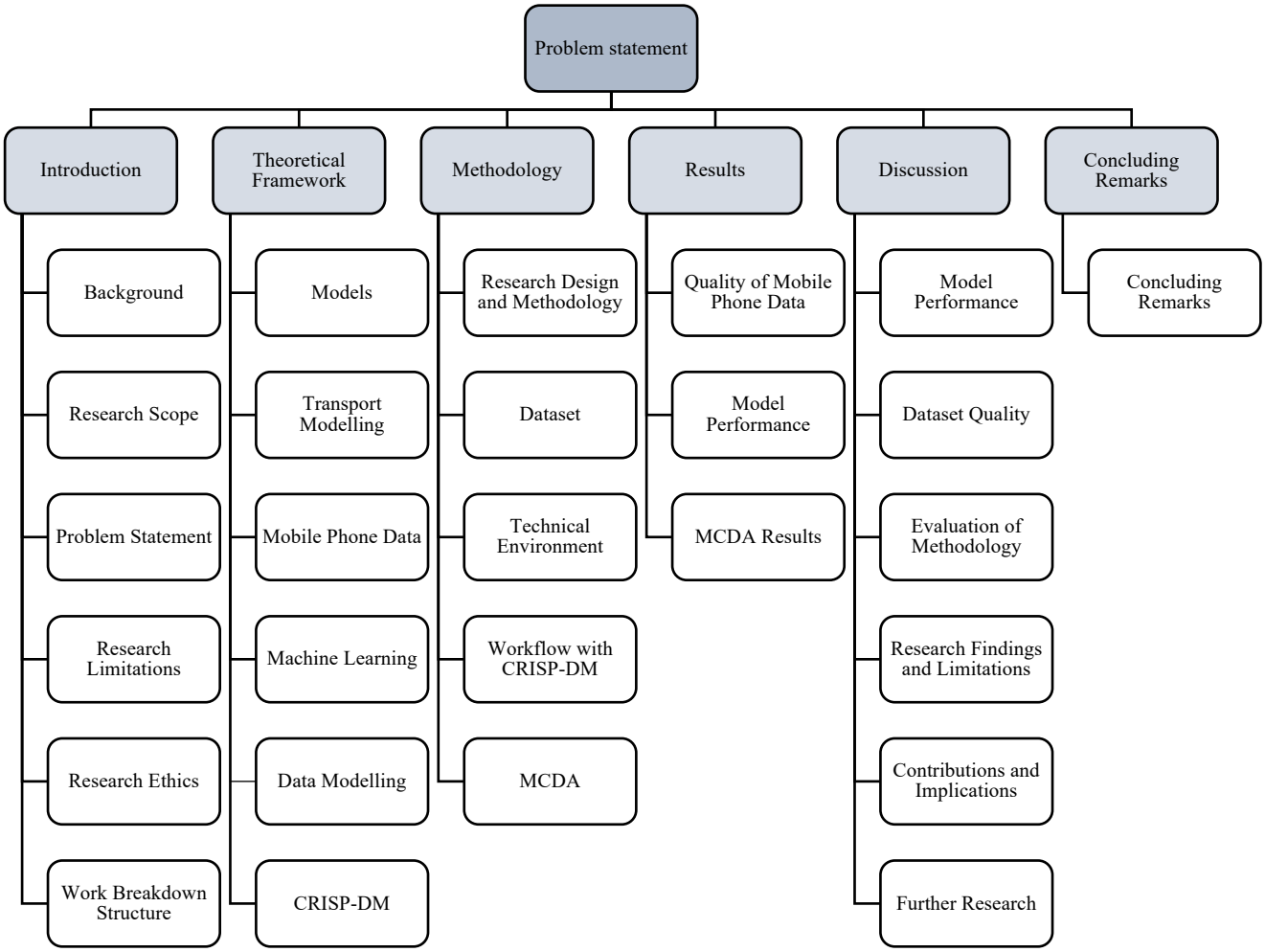


Figure 1: Illustration of the Work Breakdown Structure of the thesis.

1.6 Research Ethics

All research should adhere to high ethical standards to maintain scientific integrity and professionalism in academic work. This includes responsibilities such as authorship and accuracy of produced work. Particularly, guidelines and frameworks for safeguarding General Data Protection Regulation (GDPR) are essential (Johannessen, Christoffersen and Tufte, 2020). The safeguarding of GDPR is explained in section 0 concerning the dataset utilized in this study. Another research ethic to consider lies in the use of artificial intelligence in research contexts (NENT, 2019).

1.6.1 Usage of Large Language models

Large Language Models (LLMs) have recently become accessible to the public and are by many viewed as a concern for plagiarism and copyright issues. The Norwegian University of Life Sciences (NMBU) has developed a set of guidelines for applications of AI in a master thesis. These guidelines mandate that the thesis includes a subchapter on the use of AI, detailing how the AI was employed and verification of the information generated (Indahl and Mintorovitch, 2024).

In this thesis, the LLM by OpenAI named ChatGPT 4, was used as a source of inspiration (OpenAI, 2024a, 2024b). The model's contributions were primarily related to improving writing, formulations, and structuring paragraphs. All output was subsequently reexamined to verify no unintended information was introduced. For the data modelling in this research, ChatGPT served as a support tool for troubleshooting and suggesting next steps for coding. The following framework was mostly used for prompts to the LLM; Setting a Point of View (POV), setting a context, setting a result indicator and setting a format for answer. The POV is used as an indicator to the LLM what stance and perspective it is to take. Further, the context described the situation that was under research before setting a result indicator. An example of a prompt in this setup is presented under (*POV (1), context (2), result indicator (3), format (4)):*

"You are a data scientist (1). The goal is to investigate if this type of mobile phone data is suitable for a machine learning algorithm, that will provide value in terms of better prediction for destination choice in transport modelling (2). Analyze the dataset on what machine learning models that might be suitable for this type of data (3). Provide the answer in a table format where the column names are; name of model, description of model and how to use. (4)"

Chapter 2

Theoretical Framework

The theoretical framework introduces the relevant knowledge upon which the research of this thesis is based on. The chapter presents theories related to models, transport modelling, methodology and domain-specific theories.

2.1 Introduction to Models

The term “model” is a broad term with many ambiguities, ranging from art pieces to type of cars. A model is defined by Ortuzár and Willumsen (1994) as a simplified representation of the real world, that focuses on specific elements in a system of interest important for analysis from a particular perspective. This simplification facilitates a deeper understanding of the mechanisms of relevant variables and their associated risks, providing valuable insights. Models are utilized not only to predict future outcomes but also support decision-making processes in planning. Figure 2 presents a selection of characteristics of a model.

Simplification	• Making a representation of real-world systems to easier analyze systems.
Prediction	• Prediction of future states and outcomes
Understanding Mechanisms	• With the possibility to isulate variables models can show mechanisms behind observed phenomena.
Decision Support	• Models improve a basis of making informed decisions
Communication	• The representation of a model can make a consensus in a team presenting complex systems in an understandable form.
Hypothesis testing	• By changing variables and construction of model one can simulate under different conditions according to a hypothesis.

Figure 2 A selection of characteristics of a model, made with compiled input from (Ortuzár and Willumsen, 1994; Oxford English Dictionary, 2002; Gilbert, 2004; Valk, Driel and Vos, 2007)

Models can be tailored towards specific frameworks or domains, including physical, mathematical, conceptual, or computational models. The objective is to gain valuable insights in complex problems, while acknowledging inherent simplifications. Despite their utility, models have limitations that must be recognized and clearly defined. Generalizations, limitations and assumptions are factors that may not be true under all conditions (Ortuzár and Willumsen, 1994; Valk, Driel and Vos, 2007). These considerations are all factors building the foundation of models when looking further on more specialized topics, such as transport modelling.

2.2 Transport Modelling

In the context of new infrastructure development, transport models are extensively used in the phase of planning. Serving multiple functions such as impact assessment, decision-support, and predicting travel demand for infrastructure projects. Transport models is used as a tool by urban planners and civil engineers. Where, to be noted, transport modelling is only a segment of the whole transport planning process, but further enclose needs assessment and development of alternatives is included (Ortuzár and Willumsen, 1994). This emphasizes that developing infrastructure is a multi-dimensional problem with a variety of elements that must be addressed to make the best decision. The industry uses the transport models to solve one of those dimensions.

2.2.1 Usage of Transport Models

Understanding the underlying reasons for transportation within a community is needed when working with transport modelling. In most cases transportation is derived and is not a goal in itself, but used as a means to fulfill other needs such as work, appointments, vacation, moving goods to customer etc. This relates to a part of the characteristics of modeling transport demand. Transport demand is difficult to predict, due to complexity and variation where the purpose of transportations varies based on regular workdays, holiday, and spontaneous needs (Ortuzár and Willumsen, 1994).

Regarding transport supply, Ortuzár and Willumsen (1994) underscores that transport supply is considered a service not a good. This distinction means that transport cannot be stored and must be used where it is installed. The transport system relies on fixed assets (like roads) and mobile assets (such as vehicles) to be fully functional. Fixed assets are immovable and expensive to construct, thus it is resource-efficient to accurately estimate transport demand before developing the transport supply. Transport modelling must account for the variability of human behavior, influenced by seasonal changes due to holidays or special events (Ortuzár and Willumsen, 1994). A common method for estimating transport demand involves the four-step model.

2.2.2 The Four-Step Transport Model

The four-step transport model (4SM) is commonly used as a systematic approach for estimating future scenarios within transportation. Investing in transport infrastructure involves high incremental costs, assessing the need for construction of projects to align with actual demand. However, predicting future demand is challenging given that the average lifespan for a road project is approximately 25 years (European Commission, 2008). The four-step model is in many terms stated as the universal way of modelling in the transportation domain. The four steps of a transport model are trip generation, trip distribution, mode choice and route assignment (Andersson, Winslott Hiselius and Adell, 2018). Broad understanding of transport demand and transport supply, as explained in chapter 2.2.1, is required effectively use the 4SM.

A typical method for data collection in transport models involves conducting travel surveys. The survey is collecting data on the different elements of a trip (e.g. origin, destination, mode choice and duration) and characteristics of the individual traveler, such as age, sex, and occupation. Traditionally, this data has been collected through interviews or online questionnaires (Svaboe, 2024). Acting as the foundational dataset for the four steps in the model.

Trip generation

This initial step involves predicting the total number of trips within a specified area. A good prediction requires understanding of factors like population, employment rates, and socio-economic factors. For a further categorical division of the trips, they can be classified into specific classes based on the reason for the trip, *e.g.* for work or recreational reasons (Ortuzár and Willumsen, 1994). The total number of trips generated serves as the basis for modeling the distribution of trips across different zones within the area.

Trip distribution

Trip distribution focuses to understand and model the start point, end point, and frequency of trips (Ortuzár and Willumsen, 1994). To achieve detailed insights in travel patterns and routes, the area of interest is typically divided into zones. The zones can be sectioned based on postal numbers, a grid system or based on the zones from where the data is collected from (Ortuzár and Willumsen, 1994). In Norway, zone sectioning is standardized by the public organization Statistic Norway (Bloch, 2024). Trip distribution can be visualized, as shown in Table 1, with a two-dimensional matrix consisting of the zones in both rows and column, making an overview of the origin-destination (OD) pairs. Where T_{ij} represents the number of trips for each pair.

Table 1 General form of an OD trip matrix, inspired by Ortuzár and Willumsen (1994).

Zones	**1**	**2**	**3**	**…**	***j***	**…**	***z***
1	T_{11}	T_{12}	T_{13}	…	T_{1j}	…	T_{1z}
2	T_{21}	T_{22}	T_{23}	…	T_{2j}	…	T_{2z}
3	T_{31}	T_{32}	T_{33}	…	T_{3j}	…	T_{3z}
…							
i	T_{i1}	T_{i2}	T_{i3}	…	T_{ij}	…	T_{iz}
…							
z	T_{z1}	T_{z2}	T_{z3}	…	T_{zj}	…	T_{zz}

Trip distribution has normally been predicted using the concept of *generalized cost.* The formula for generalized cost takes into consideration a set of factors influencing travel choices. It incorporates monetary expenses, time spent traveling, convenience or discomfort of the journey and other non-monetary factors that affect individuals' transportation decisions (Ortuzár and Willumsen, 1994).

The formula for generalized cost is given by:

$$generealized\ cost = a_1 t_{ij}^{v} + a_2 t_{ij}^{w} + a_3 t_{ij}^{t} + a_4 t_{nij} + a_5 F_{ij} + a_6 \phi_j + \delta \qquad (2.1)$$

Where,

t_{ij}^{v} is the in-vehicle travel time between *i* and *j;*

t_{ij}^{w} is the walking time to and from stops (stations);

t_{ij}^{t} is the waiting time at stops;

t_{ij}^{v} is the interchange time, if any;

F_{iJ} is the fair charged tot travel between *i* and *j*

ϕ_j is a terminal (typically parking) cost associated with the journey from *i* to *j*;

δ is a *modal penalty*, a parameter representing all other attributes not included in the generalized measure so far, e.g. safety, comfort and convenience;

$a_{1\ldots6}$ are weights attached to each element of cost; they have dimensions appropriate for conversion of all attributes to common units, e.g. money or time

(Ortuzár and Willumsen, 1994)

Mode choice

Mode choice is the phase in the four-step transport model where travelers select their mode of transportation. The mode includes everything from walking to public transport. This choice is influenced by variety of factors including cost, time, convenience, and personal preferences. Behavioral aspects, which can be challenging to predict, alongside simplifications is often required to represent the model for mode choice. The mode decision is derived from a distribution modelled based on data from transport surveys (Ortuzár and Willumsen, 1994).

Route assignment

Route assignment is the final step of the four-step model and involves determining the specific routes that vehicles or people take during their trips. This process involves allocating trips within the transport network and identifying the path of least resistance by considering variables such as travel time, congestion, tolls etc. Then assuming the individual will choose the path of least resistance. Effective route assignment facilitates more informed decision-making in infrastructure projects and aids in the simulation of future traffic scenarios (Ortuzár and Willumsen, 1994).

2.2.3 Discrete Choice Models

Discrete choice models (DCMs) are used in transport modelling for representing the decision-making process for particularly destination choice and mode choice. These models facilitate an understanding of how individuals decide between various options for modes of transport or destinations. A key concept of DCM is *utility*, which is defined as perceived benefit or satisfaction an individual aims to maximize from a choice. The utility associated with an option is derived from its belonging characteristics (Ortuzár and Willumsen, 1994). Multinomial logit model is commonly used for choice models. The multinomial logit model estimates the probability for choosing a particular destination or transport mode, taking into consideration both the characteristics of the option and the socio-economic attributes of the traveler, typically collected from a travel survey. This model allows for the simulation of how shifts in characteristics will influence mode and destination choice (Ortuzár and Willumsen, 1994). Discrete choice models are now increasingly incorporating alternative data sources, many in the form of Big Data, such as mobile phone data, to enhance their accuracy and applicability.

2.3 Mobile Phone Data

Mobile phones are a significant source of a variety of data, collectively reffered to as mobile phone data (MPD). This data is derived from radio frequency signals that can track the location of phones using cellular networks, Wi-Fi, GPS and Bluetooth (Zhenzhen Wang, 2017). These signals are received by cellular communication antennas placed throughout various regions based on the local demand for service (Anda, Erath and Fourie, 2017).

Telecommunications operators can provide three types of data: Call Detail Records (CDR), Passive Signaling Data (PSD) and Active Signaling Data (ASD). Each type differs primarily in the frequency of data recording. CDR data is generated by calls or messages sent from the phone. PSD signal occurs when the phone connects to a network, and ASD is generated when an external request pings the phone to update its location (UN council of Experts in Big Data, 2023). Contrary to what might be expected, PSD is considered active communication due to the continuous connectivity to the network, whereas ASD is referred as passive communication because location updates occur less frequently (Goves and Hemmings, 2016). Anda, Erath and Fourie (2017) note that phone tracking can be triggered by either a network connection or an event, as detailed in Table 2, where the phone updates its location.

The tracking of mobile phones leads to a large number of observations of the location of people, which can be taken advantage of in transport modelling.

Table 2 Signal triggers and types for tracking location of mobile phones. Made with information from Anda, Erath and Fourie (2017) and UN council of Experts in Big Data (2023)

Signal trigger	**Signal type**
Connects to cellular network	PSD
Occurring call and move between two cell areas	CDR
On standby and moves into a new Location Area (LA)	ASD
A periodic location update (a ping about every 2 h)	ASD
The phone is placing or receiving a call	CDR
When a Short Message Service (SMS) is used (both sending and receiving)	CDR
When the phone connects to the internet	PSD

2.3.1 Mobile Phone Data in Transport Modelling

Mobile phone data provides a relatively new data source in transport modelling, where the main strength lies in the large number of observations provided, unlike traditional travel surveys. For instance, in Tønsberg, with a population of 55 387 people (as of 1st of January 2023) (Thorsnæs *et al.*, 2024), a typical travel survey with a 1 % response rate would yield only 554 responses (Anda, Erath and Fourie, 2017). Considering the geographic division, the zone sectioning is based on the location of telecommunications towers, which capture and interpret signals from mobile phones. This type of sectioning has its strengths and limitations. One significant challenge involves data processing errors related to signal reception by the towers, particularly at the borders of zones, which can lead to misclassification (Nilsen, Uteng and Myrberg, 2021). Zones can vary in size, from as small as 500x500 m in urban areas to as large as 7,5x7,5 km in rural areas. Introducing considerable uncertainty on the precise location of a mobile phone (Goves and Hemmings, 2016).

As noted in section 2.3 the registration of the location of mobile phone data involves signals emitted by the phone and received by a telecommunication tower. The signals can be categorized as passive or active communication; passive when the phone is not in use and active when the phone is using actively used. Active communication results in more frequent logging of the phone's location and time. Figure 3 illustrates the differences between PSD and ASD signaling, showcasing that the level of data detail also depends on the type of signal received.

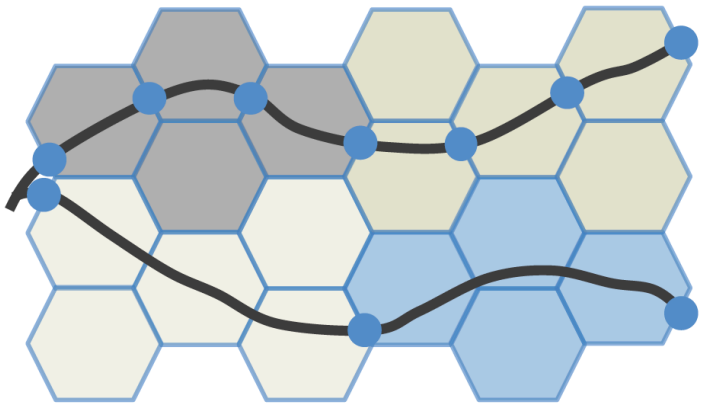


Figure 3 The difference between PSD (top line) and ASD (bottom line). Figure from Goves and Hemmings (2016)

A concept in utilizing mobile phone data for transport modelling is the definition of a “trip”. In general, a trip is every movement outside the house, workplace, or school. The trip is independent on the distance, duration or means of transport (Nilsen, Uteng and Myrberg, 2021). Within the context of transport modelling using mobile phone data, a trip is specifically defined as an instance where a mobile phone crosses a zone border (Goves and Hemmings, 2016). This crossing is registered when the signal from a phone is detected in a different zone than previously recorded. A zone is considered as the endpoint of a trip if the device remains within it for a specified duration, often referred to as “dwell time”. In this study, using data from Telia, the dwell time is set at 20 minutes (Wismans *et al.*, 2018).

The limitations of the mobile phone data primarily include its inability to directly provide insight in modal choice or the route choice. Additionally, when a telecommunication tower covers multiple zones, assumptions must be made regarding the exact location of the phone. The data also includes timestamps that enable the differentiation of travel patterns across various times, such as weekdays versus weekends or mornings versus evenings, facilitating the analysis of trends and variations in travel behaviors (Anda, Erath and Fourie, 2017; Nilsen, Uteng and Myrberg, 2021).

Compared to traditional travel surveys, mobile phone data offers a larger number of respondents and better geographical coverage, making it more dynamic. The survey gives more information about travel patterns (where) and travel behavior (how and why). The survey also gives more in-depth information about the traveler, whereas mobile data is completely objective and only gives travel start- and endpoint (UN council of Experts in Big Data, 2023; Svaboe, 2024)

Given the vast volume of data involved in mobile phone data, it is of interest to see how it can be utilized in a context which leverages the possibilities with artificial intelligence and machine learning. These technologies can potentially enhance the analysis and utility of transport models.

2.4 Machine Learning

Machine learning (ML), a subset of artificial intelligence, use algorithms grounded in statistical and mathematical models to complete given tasks. Rather than having humans to calculate and follow calculation rules, the task is transferred to the computer. The algorithms progressively improve their performance by analyzing feedback on their outputs. Machine learning is commonly divided into two main types: unsupervised learning and supervised learning. Unsupervised learning involves training the algorithm without providing predetermined answers, allowing the machine to find patterns in the data. Conversely, supervised learning involves providing known answers to the algorithm, enabling it to test its performance and adjust based on this information (Raschka & Mirjalili, 2017). These two approaches have distinct characteristics and applications, where this section gives a more detailed explanation.

2.4.1 Learning Techniques in Machine Learning

As the field of machine learning continues to evolve, numerous algorithms have been developed. However, they generally adhere to the same principles and learning techniques. Rashhcka & Mirjalili (2017) separates the different techniques by supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning

In Supervised learning, algorithms are trained using labeled data, which provides explicit feedback and predicted outcomes. This method involves the algorithm making predictions and adjusting its response based on whether the outcomes are correct, effectively learning from its errors to enhance accuracy. For supervised learning, a dataset with a known response variable is necessary, which allows for the classification of observations into distinct classes based on features. The features are used by the algorithm to find the patterns and establish functions for accurate classification (VanderPlas, 2016; Raschka and Mirjalili, 2017).

Table 3 Overview of how supervised learning is split into a data frame of features and a data frame of response variable. Made with inspiration from VanderPlas (2016)

Observation	**Variable 1**	**Variable 2**	**. . .**	**Variable m**
1	X_{11}	X_{12}	…	X_{1m}
2	X_{21}	X_{22}	…	X_{2m}
. . .	. . .	. . .	. . .	. . .
n	X_{n1}	X_{n2}	…	X_{nm}

Observation	**Response**
1	y_1
2	y_2
. . .	. . .
n	y_n

For supervised learning, the dataset used is divided into two parts: a training set and a testing set. The division ratio between these sets can vary depending on the size of the dataset, with splits ranging from 70/30 to 99/1 for training and testing, respectively. A common split is 80% for training and 20% for testing. Supervised learning excels when there is a need to use human resources to classify an observation, because by training the algorithm based on the empirical data the computer will be able to classify on its own (Raschka and Mirjalili, 2017).

Unsupervised learning

Unsupervised learning is used when there is no designated response variable within the dataset. It differs from supervised learning in that the correct answer is not predefined before training the model. The primary goal is to discover underlying structures within the data by identifying patterns and extracting information without explicit labels. This type of learning is suitable when the data lacks clear categorization. The algorithm rather tries to find patterns or similarities between observations by for instance clustering them. The patterns can be proximity to cluster centers, where the clusters then represent the different classes or groups within the data (Raschka and Mirjalili, 2017;VanderPlas, 2016).

Reinforcement learning

Reinforcement learning is a decision-making process that includes a reward system. The system is allowed to enhance its performance through interaction with its environment. Unlike supervised learning, the feedback in reinforcement learning does not come from a ground truth label but rather from a reward function. The learning process involves maximizing rewards through a trial-and-error method (Raschka and Mirjalili, 2017). Raschka and Mirjalili (2017) compares it to a game of chess, where each creates a new scenario in the game. The algorithm then strives to determine a sequence of moves that will lead to victory, guided by the positive or negative rewards received after each move.

Each of these learning techniques offers distinct methodologies for constructing models that best fit the available data and yield relevant results. The following section presents the underlying theory for the selection of algorithms tested in this research.

2.4.2 Classification Algorithms

Classification algorithms aim to predict the category to which a new observation belongs by learning from a dataset in the training phase. There are different approaches possible for solving a multiclass classification problem. The common denominator is that they use data to train itself for better predicting. When the goal of this research is to predict the destination choice and based on the given data, it is most suitable to choose from a selection of supervised learning algorithms. For this study it is used four different types of algorithms which each has its own way of classifying data samples.

Logistic Regression

Logistic Regression is a widely used classification algorithm, known for its simplicity and ability to perform well on linearly separable data. Primarily a binary classification algorithm, it can be extended to multiclass classification with the use of One-vs-Rest (OvR) technique. The logit function is used for calculation of the linear relationship among the value of features and the natural logarithm of the odds ratio. Where the odds ratio is the favorable odds for a given event (Raschka and Mirjalili, 2017).

$$logit(\, p(\, y = 1 \mid \boldsymbol{x})) = \, w_0 x_0 + \, w_1 x_1 + \cdots + \, w_{\boldsymbol{0}} x_0 = \sum_{i=0}^{m} w_{\boldsymbol{i}} x_i = \boldsymbol{w}^T \boldsymbol{x} \qquad (2.2)$$

Where $p(\, y = 1 \mid \boldsymbol{x})$ is the probability for the sample to belong class 1 given **x** features, and w_i is the weighting of the feature *I* (Raschka and Mirjalili, 2017). Where the Logistic Regression algorithm further calculates the probability that an observation belongs to a given class, calculated with the logistic sigmoid function.

$$\phi(z) = \frac{1}{1 + \, e^{-z}} \qquad (2.3)$$

Where z = $\mathbf{w}^{\mathrm{T}}\mathbf{x}$, indicating that the sigmoid function is the inverse of the logit function. (Raschka and Mirjalili, 2017)

Support Vector Machine

Support Vector Machine (SVM) focuses on maximizing the margin, defined as the length between the decision boundaries (hyperplanes) that separate different classes, and the data samples that are the nearest to this hyperplane (Raschka and Mirjalili, 2017). This characteristic makes SVM robust to outliers as it primarily trains on data points nearest to the hyperplane, as illustrated in Figure 4. VanderPlas (2016) states its power and flexibility, making it suitable for a variety of complex classification problems.

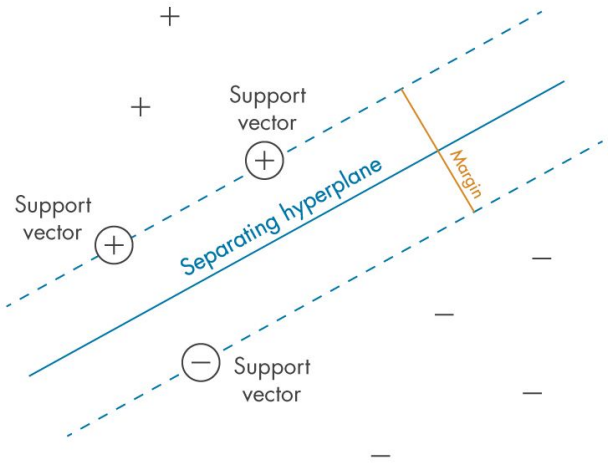


Figure 4 Illustration of separation with hyperplane and support vectors. Illustration from MathWorks (n.d.).

Gaussian Naïve Bayes

Gaussian Naïve Bayes (GaussianNB) is a simple generative model, where its strength is being very fast and can work well for high-dimensional datasets. It is a generative model due to the specification of the hypothetical random process that has generated the data. The algorithm uses Bayes's theorem to compute the probability of a label based on the observed features.

$$P(L \mid features) = \frac{P(\ features \mid L) * P(L)}{P(features)} \tag{2.4}$$

Where, *(P (L | features)* is the probability of a label given the observed features. *P(features | L)* is the probability of a feature given a label. *P(L)* and *P(features)* is respectively the general probability of a label and a feature.

The model works under the assumption that each label from the data is taken from a given distribution, in this case a Gaussian distribution. This leads to the model being simple and have no hyperparameters to tune, but due to its simplicity it is ideal as a baseline classifier (VanderPlas, 2016).

Random Forest

Random Forest (RF) is an ensemble technique that uses multiple decision trees to perform regression and classification tasks. VanderPlas (2016) states it as one of the most powerful algorithms due to its ease of use and strength for regression and classification problems. Especially the algorithm can handle variance in the dataset and large amounts of data. The algorithm involves creating multiple trees from bootstrap samples of the training data, choosing a subset of features at each node to split based on the objective function. These steps are repeated a set number of times, where the higher the number the better performance, but at a larger computational cost (Raschka and Mirjalili, 2017). The split is normally done on the feature which leads to the largest information gain (IG), aiming for purity in the leaves of the tree. This approach can lead to overfitting, making a need for a pruning constraint which limits the algorithm on the depth of the trees (Raschka and Mirjalili, 2017). Figure 5 demonstrates the decision process of Random Forest in addressing a multiclass problem, such as selecting a transportation mode for commuting.

For the algorithms to perform as good as possible on the given task the underlying settings, called hyperparameters, needs to be adjusted so that it can manage the training with the provided data.

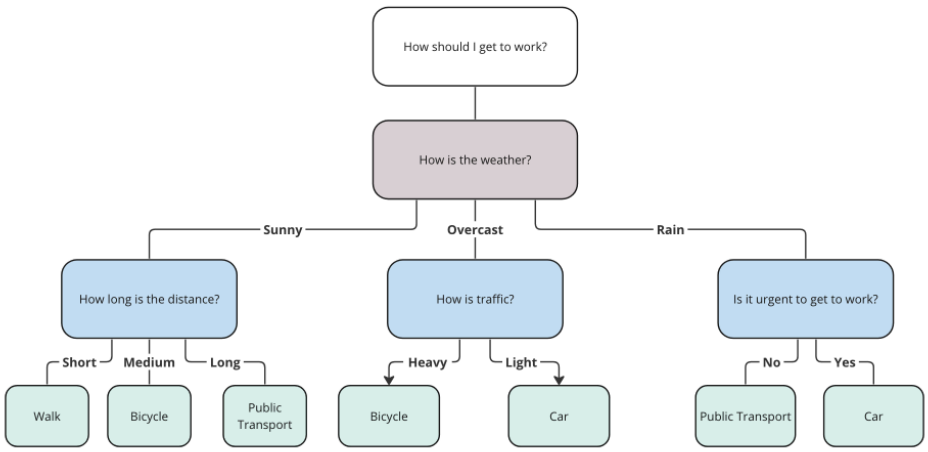


Figure 5: An illustration of how a decision tree works in a Random Froest. Drawn with inspiration by VanderPlas (2016)

2.4.3 Hyperparameter Tuning

In machine learning and artificial intelligence, hyperparameters play an important role in the architecture of models to gain the best performance (Owen, 2022). Unlike parameters, which lears automatically from the data during training, hyperparameters are set before the training process begins. Hyperparameters can be seen as the external factors used to adjust a machine learning algorithm. Examples of configurable hyperparameters are learning rate, depth, number of trees, number of layers and number of neurons. These settings effect the model's ability to generalize to unseen data. As illustrated in Figure 6, a machine learning model can have different levels of detail for classification. A model is not performing well if it is either overfitted or underfitted. Thus, fine-tuning hyperparameters are a tool to adjust the fit, leading to enhancement of the performance and efficiency of machine learning models (Raschka and Mirjalili, 2017; Owen, 2022; MathWorks, 2024).

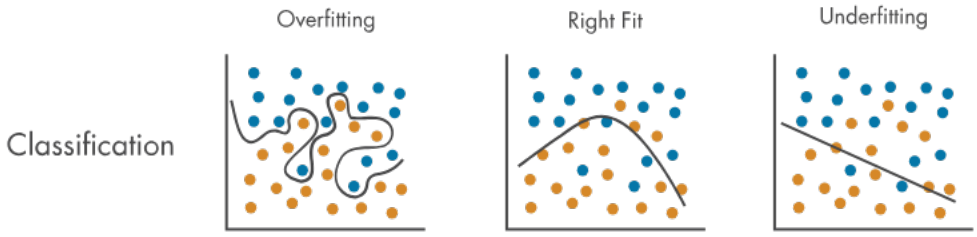


Figure 6 Illustration on how overfitting, optimal fitting and underfitting for a classification problem. Figure from MathWorks (2024)

To determine the optimal set of hyperparameters, techniques such as grid-based search and Bayesian optimization are commonly employed. Grid-search involves a predefined searching through a manually specified subset of hyperparameters, whereas Bayesian optimization provides a more dynamic approach, having a built-in function for adjusting its hyperparameters as the model is training. The challenge in hyperparameter tuning lies in the often complex, non-linear interdependencies between different hyperparameters and the substantial computational resources required for extensive experimentation (Raschka and Mirjalili, 2017; Frazier, 2018; MathWorks, 2024).

2.5 Cross Industry Standard Process for Data Mining

The Cross Industry Standard Process for Data Mining (CRISP-DM), as defined by Chapman et al. (2000), outlines a structured approach for developing data-driven models across various industries. It serves as both a project management and data analysis framework. CRISP-DM is composed of six sequential phases designed to guide the order in which tasks should be executed. These phases are: *Business understanding, data understanding, data preparation, modelling, evaluation* and *deployment.*

As visualized in Figure 7, the process is represented as a cyclical diagram, emphasizing its iterative nature. If a developed model fails to meet business criteria, the process mandates a return to the Business Understanding phase for further iteration until the model is deemed satisfactory for deployment. Iterations may also occur between business understanding and data understanding, data preparation and modelling. These iterations allow for continual refinement of the business insights and data strategies as limitations are encountered or new information is acquired throughout the project lifecycle (Chapman et al., 2000).

It is further given an outline of what activities and focus to be included in each of the steps for striving towards a standardized process in the data mining.

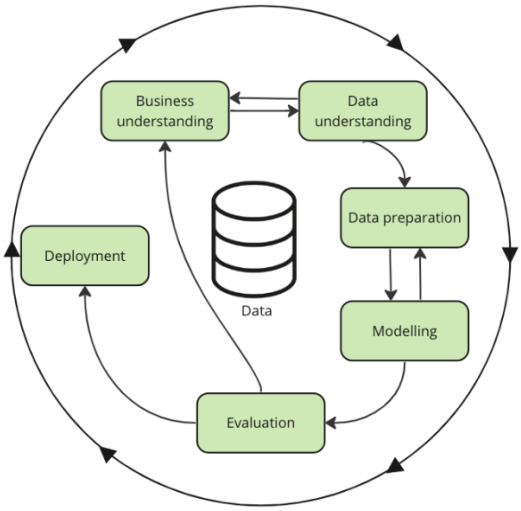


Figure 7 Cross-Industry Standard Process for Data Mining. Figure made with inspiration from Chapman et al. (2000)

Business understanding

Business understanding involves understanding the business objectives and the challenges that the industry seeks to address. This phase serves as an assessment of the current state of the industry, identifying the technological level at which it operates. Furthermore, it facilitates the recognition of project limitations and challenges (Chapman, et al., 2000). The business understanding should be approached from both a macro and a micro perspective. At the macro level, the superior business goals of the industry provide the foundational basis for initiating the project. Conversely, at the micro level, specific objectives for individual departments, teams, products, or services are identified. This is a fundamental step to complete thoroughly before embarking on the data understanding (Chapman, et al., 2000).

Data understanding

Data understanding involves getting an overview of the collected and available data. Further, building understanding of what the dataset contains and become familiar with it. Evaluating data quality and information of interest is here important for further development. This leads to first insights in the potential applications of the data in alignment with the business objectives, before proceeding with the data preparation (Chapman, et al., 2000).

Data preparation

Data preparation involves manipulating the dataset to ensure its suitability for modelling. It is important to perform statistical analysis to appropriately manage the data. It includes cleaning and formatting the data to eliminate inconsistencies, such as addressing missing values, converting data into the correct format, and managing outliers. Additionally, it may involve creating derived attributes or merging datasets, provided that an initial understanding of the second dataset has been established. Typically, all irrelevant information is removed from the dataset to make the model more computational efficient and interpretable for the modelling (Chapman, et al., 2000).

Modelling

A data model simulates the real world by representing objects and their interactions within a computer system. These models are used to analyze how systems respond to external changes or for extracting and visualizing insights from data sources (Rossen, 2021). Classification algorithms, as described in chapter 2.4.1, are a prominent type of algorithm used in data modelling (Raschka and Mirjalili, 2017). Modelling involves generating a data model based on the dataset, which is aiming to give relevant results for solving the business objective. Some models require a specific format on the data, and therefore it is normal to go back for further data preparation. Different techniques, models and approaches can be used and compared up against each other based on performance metric and further evaluation (Chapman, et al., 2000).

One of the main factors for an accurate data model is the tuning of hyper parameters and keeping a low bias in the model. With an optimized and tuned model, it has a chance of greater performance and generality (Claesen and De Moor, 2015). As Figure 8 is illustrating there is an optimal model complexity. With a low model complexity it has a high bias, indicating that there is an oversimplification and underfitting in the model. It is not complex enough to find patterns in the data, leading to a high error. On the other side, a model with high complexity will gain high variance which leads to much error. A model with high variance will perform poorly on unseen data. The goal is to have enough complexity to find the patterns, but not too much, so that the model can perform on unseen data as well (Fortmann-Roe, 2012).

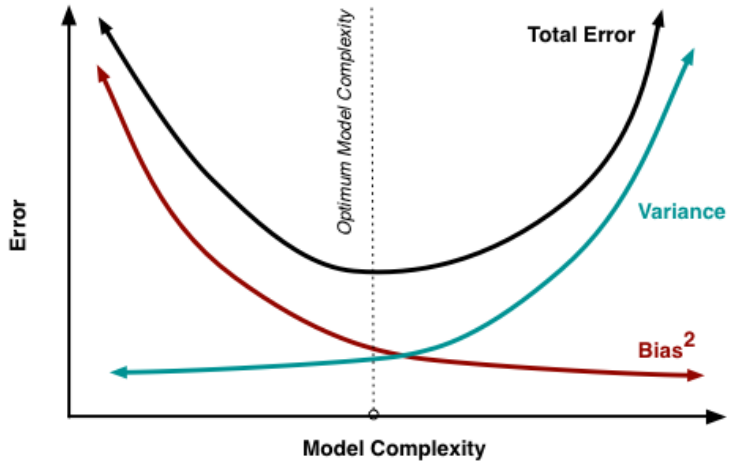


Figure 8: bias-variance trade-off. When a model has an increase in complexity it also increases variance. Figure from Fortmann-Roe (2012)

Evaluation

Evaluation involves evaluating the different models, both up against each other, but most importantly, evaluated up against the problem which is trying to be solved. To further address in what degree the model brings value to the industry related to the business objectives. A key element is to determine if some objective is not sufficiently covered for the model to be ready for deployment (Chapman, et al., 2000).

Deployment

Deployment is the step where the model is to be implemented in a real-time decision-making environment, new technology implementation can always be a challenge, and it is important to consider how the model is organized and presented for the onboarding of the users. The users need an interface for handling the model, which could be anything between a report, website, or software-plugin, depending on the needs. Often, it is not the data scientist being the direct user of the model, and therefore the degree of technical skills needed to operate the model should be limited (Chapman, et al., 2000).

Chapter 3

Methodology

In this chapter, the methodical approach for addressing the problem statement and research question is described. The approach applies the CRISP-DM framework presented in section 2.5. The chapter articulates the rationale behind the chosen research design, presents an overview of the mobile phone dataset, describes the workflow, details the steps in the data mining, and lastly describes the method for evaluation of results.

3.1 Choice of Research Design and Methodology

Collaborating with a company provided access to expert support and a comprehensive dataset, which influenced the choice of research design to lean more towards an industry near approach. A goal was for the approach to connect with the potential value it can provide for the industry. The research design was chosen to align with the defined problem statement and research questions. This alignment impacts data collection and the analytical approach. With the large dataset to be analyzed it was latent to have a quantitative approach.

Consideration of the research's reliability and validity is important. For the research to be valid it needs to have transfer value into the industry or be potential for further research. Additionally, the methodology should maintain reliability, ensuring that it builds trust and allows for reproducibility by others. The selection of the CRISP-DM framework was predicated on its ability to enhance cost-efficiency, standardization, speed and reliability in data mining projects (Wirth and Hipp, 2000). CRISP-DM was preferred over other frameworks such as KDD and SEMMA due to its comprehensive documentation and holistic approach (Azevedo and Santos, 2008).

Newer frameworks specializing in machine learning and artificial intelligence, such as BizML, have been proposed by Siegel (2024). However, it falls for CRISP-DM for the same argument of not being well enough documented and tested. An expanded method of the CRISP-DM is Cross-Industry Standard Process for Machine Learning with Quality assurance (CRISP-ML (Q)). Studer et al. (2021) states that CRISP-DM has short-comings on two points: Firstly, it does not cover how machine learning models are to be used in real-time for decision making in a more operational setting. Over time machine learning models will degrade, implying there is a need for a guidance for updating and maintenance of ML models. Secondly, CRISP-DM fails to address quality assurance where the focus is to find risks related to adapting it into business, as well as increasing interpretability of a machine learning model. This research is limited to not to focus

on the deployment of the model, the first argument falls short. The second argument is more relevant but would be a to large increase in the scope of the research for it to be thoroughly investigated and evaluated. Since CRISP-ML is derived from CRISP-DM they are very similar in their build-up, and can take some inspiration from CRISP-ML(Q). Thus, CRISP-DM remains the preferred framework for this study.

3.2 Dataset

The data used in the machine learning models origins from two different sources: mobile phone data and zone information. This section provides an overview of the data collection and the characteristics of the dataset.

3.2.1 Data Collection

The primary dataset has been a dataset acquired from mobile phones. The data is collected from an area under interest in Tønsberg municipality, Norway, as marked in red in Figure 9. Norway is divided into about 14 000 zones, whereas the area of interest in Tønsberg municipality covers 589 of them (Bloch, 2024). The data was gathered in 2019 by the Norwegian teleoperator Telia through their service in crowd insight for transportation (Telia, 2024a). Data points were generated through interactions with telecommunication towers within the specified area, as detailed in chapter 2.3. This implies that other types of mobile phone data, such as GPS, is neglected in further mentioning of mobile phone data. Rambøll acquired the data and preprocessed it into fitting their standards for zones and feature-naming and merged it with the zone information.

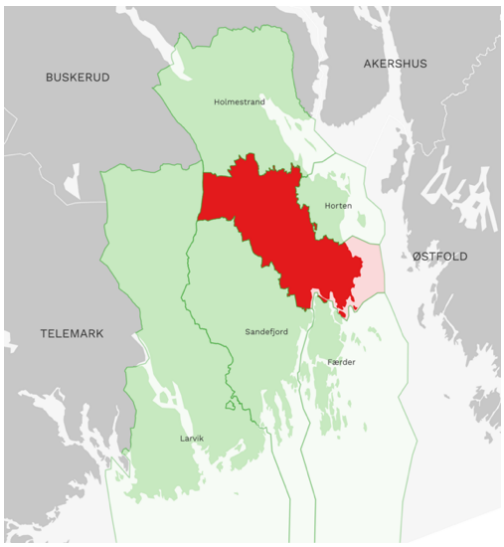


Figure 9 Tønsber municipality. Figure from Store Norske Leksikon (2024).

3.2.2 General Data Protection Regulation

Given that the data collection involves tracking individual locations, compliance with the General Data Protection Regulation (GDPR) is imperative. The responsibility for ensuring GDPR compliance for data collection primarily lies with Telia. To safeguard privacy, the person is assigned a random ID number to each phone at the start of the day. The location of the phone is then monitored at intervals on occurring events as described in chapter 2.3. The travel data is recorded, and the ID is deleted at the end of the day, before a new ID is issued the following day. For more information see the privacy guidelines of Telia (2024b). Further processing with the data complies with The Norwegian Personal Data Act (2018).

3.2.3 About the Data

The collection of data extended over a full year, resulting in a total of 79 256 205 trips being registered, averaging 1 444 registered trip per resident of Tønsberg, which is in stark contrast to travel surveys (Anda, Erath and Fourie, 2017; Thorsnæs *et al.*, 2024). This leads to a satisfactory representation of the population in the area of interest. The dataset is structured and consist of numerical values. In addition to the mobile phone data, a supporting dataset has been necessary to merge with the pre-processed dataset. This dataset contains information about the different zones, such as number of workplace types, distance between zones and population. Each zone-pair constitutes an observation in the machine learning context, with the number of trips between these zones serving as a feature. The dataset has not been applied related to research before but have earlier been used in projects in Rambøll. The dataset is applied in its full, and a brief overview of the dataset is given in Table 4.

Table 4 Overview of the dataset made available for the research in the thesis.

	Pre-processed Mobile phone data
Name	Cube-data_Tberg_total_24032021.xlsx
Source	Rambøll Norway AS, Telia AS
Period for data collection	2019
File type	Microsoft Excel spreadsheet (.xlsx)
Number of observations	162 517 zone pairs (79 256 205 Trips)
Number of different zones	589
Response variable	Destination zone

The preprocessed dataset has 16 features, with feature names presented in Norwegian and English. The features *Walk, Bicycle, Car, Carpass and PT*, seen in Table 5, originate from travel survey data. The features *Origin-zone* and *Destination-zone* are linked together making each row represent a unique origin-destination pair. This directly affects the Trips and Distance feature, as they are derived from the information gained from the zone pairs. The table further shows that all the features have numerical data.

Table 5 Overview and description of the columns in pre-processed mobile phone data set

Column	Renaming	Type of data	Description
Fra-sone	Origin-zone	Numerical (integer)	Origin-zone, the zone where the trips start.
Til-sone	Destination-zone	Numerical (integer)	Destination-zone, the zone where the trip ends.
Fra-Grk	origin-area	Numerical (integer)	Origin-zone (fra-sone) converted into basic statistical unit; a unit used by Statistics Norway.
Til-Grk	Destination-area	Numerical (integer)	Destination-zone (til-sone) converted into basic statistical unit; a unit used by Statistics Norway.
Reiser	Trips	Numerical (float)	Number of trips between the pair of zones.
Ndays	Ndays	Numerical (float)	Number of days for data collection.
Distance	Distance	Numerical (float)	Distance between the centroids of the zone pairs
TotBef	TotPop	Numerical (float)	Total population in ‘origin-zone’ (fra-sone).
Arb	Work	Numerical (integer)	Number of public workplaces in ‘destination-zone’ (til-sone).
Handel	Retail	Numerical (integer)	Number of retail workplaces in ‘destination-zone’ (til-sone).
FriHenPriv	RecDelPriv	Numerical (integer)	Number of workplaces related to recreational activities, pick-up and delivery, and other private workplaces in ‘destination-zone’ (til-sone).
Gange	Walk	Numerical (float)	Number of aggregated from travel survey data.
Sykkel	Bicycle	Numerical (float)	aggregated from travel survey data.
Bil	Car	Numerical (float)	Travels by car, aggregated from travel survey data.
Bilpass	Carpass	Numerical (float)	Car passengers, aggregated from travel survey data.
PT	PT	Numerical (float)	Public Transport trips, aggregated from travel survey data.

3.3 Technical Environment

The technical environment for data modelling can affect the results. Thus, it is here given an overview of used hardware and software. The primary hardware applied is a 2020 Apple MacBook Pro with a four-core Intel Core i5 processor, 16 GB of RAM and 256 GB SSD, running macOs Monterey 12.4. Python, version 3.9.18 (Python, 2023), is the programming language used in the research, where code has been written in Jupyter Notebook (Jupyter, n.d). Google Colaboratory (Google, n.d.) was used for additional computational power needed for data processing and modelling. A variety of libraries has been used, presented in Table 6. The code and dataset in full can be found by following the noted GitHub link[1].

Table 6 Overview of the python libraries used in the data model.

Library	Version	Description
NumPy	1.26.3	NumPy (Numerical Python) offers a variety of functions for data management. It as an easy-to-understand user interface directed towards arrays, mathematical and statistical analysis and calculations (Harris *et al.*, 2020).
Pandas	2.1.4	Pandas is a library that is commonly used for the handling of databases. It is used to convert the data into the correct format (from csv, xlsx, SQL, etc.) and store them in DataFrames. It is able to handle large amounts of data. Pandas is based on series and DataFrames which is utilized to change and structure data. A series is one-dimensional, consisting of either a row or a column, and a DataFrame is two dimensional, consisting of both rows and columns (McKinney, 2010).
Scikit-learn	1.2.2	Scikit-learn is an open-source library used for accessing machine learning algorithms. It gives access to already developed algorithms that are commonly used for data modelling. It can further be used for visualization of data, hyperparameter tuning and functions for evaluation of performance (Pedregosa *et al.*, 2011).
Matplotlib	3.8.0	Matplotlib is one of the most common libraries used for visualization. It can easily show a great variety of graphs including heatmaps, line graphs, bar charts, and multidimensional plots (Hunter, 2007).
Seaborn	0.12.2	Seaborn is also a library used for visualization. It is based on Matplotlib but give more informative statistical graphics and complex graphs with simple commands. It is tailored to work with Pandas data structures. Some common graphs are heatmaps, time series and violin plots (Waskom *et al.*, 2022).

[1] https://github.com/kristiols/olsen_master_thesis.git

3.4 Workflow with CRISP-DM

CRISP-DM serves as the methodological framework applied in this research, chosen for its applicability to industry implementations. The approach was evaluated as suitable for addressing the research questions and to be able to build a data model with the given data. This research focuses on the first five out of the six phases of CRISP-DM: *Business understanding, data understanding, data processing, data modelling* and *data evaluation*. Deployment is excluded due to the novelty of this research and choice of research scope. Given that integration of machine learning and mobile phone data is at the preliminary stage it is considered premature to delve into the deployment phase. Each phase is developed with detailed steps inspired from Corrales, Ledezma and Corrales (2015), for the methodology to have a thorough foundation of a systematic approach as shown in Figure 10.

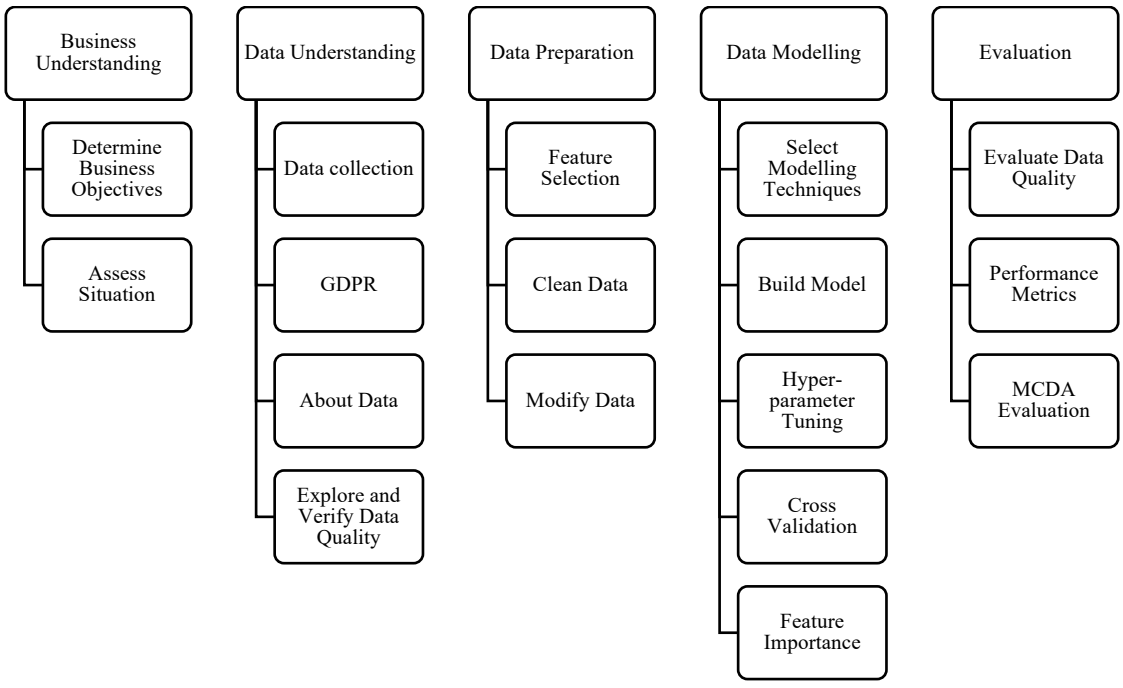


Figure 10: Illustration of workflow of methodology with the CRISP-DM framework. Inspired by Corrales, Ledezma and Corrales (2015), Chapman et al. (2000) and Wirth and Hipp (2000).

3.4.1 Business Understanding

The initial phase involves comprehending the business environment and operational dynamics within the industry to build a foundation for further construction of the data model. According to Chapman et. al. (2000) it is important to understand the business objectives is with the data-mining process. This phase has followed the steps presented in Figure 10, and has been conducted in close collaboration with Rambøll, providing insights about transport modelling and how the current operational practices. Discussions also explored hypothesis regarding the potential of the mobile phone data datasets and how AI-models can be used in the industry. This guided the development of problem statement and research questions.

Determine business objectives: Development of problem statement and research questions

Through the work with the business understanding it was addressed what types of challenges the transport modelling industry were facing, and Rambøll's thoughts on how mobile phone data can be integrated into transport models. This lead to the formulation of the problem statement and research questions presented in chapter 1.3 and chapter 1.3.1, derived from the research scope and knowledge gaps addressed in chapter 1.2. The research limitations are presented in chapter 1.4.

Assess situation: Understanding Transport modelling and mobile phone data

Assessing the situation involves building an understanding of the environment for where the data model is to be utilized. The book Transport Modelling by Ortúzar and Willumsen (1994) was used for understanding the foundational principles of transport modelling. Subsequent searches for literature were conducted on Google Scholar and Oria. The search aimed to identify where artificial intelligence has been applied in a transport modelling context, instances where mobile phone data have been used in transport modelling, and if any research combining both mobile phone data and artificial intelligence. An overview of key literature for transport- and mobile data understanding is presented in Table 7.

Table 7 Overview of key literature for transport modelling and mobile phone data

Authors	Title	Type
Ortuzár and Willumsen (1994)	Modelling Transport	Book
Anda, Erath and Fourie (2017)	Transport modelling in the age of big data	Journal Article
UN council of Experts in Big Data (2023)	Use of Mobile Phone Data in Transportation	Report
Caceres, Romero and Benitez (2020)	Exploring strengths and weaknesses of mobility inference from mobile phone data vs travel surveys	Journal article
Essadeq and Janik (2021)	Use of Mobile Telecommunication Data in Transport Modelling – A French Case Study	Journal Article

3.4.2 Data Understanding

The data understanding phase aims to discern valuable insights from the dataset that can influence decision-making in transport modelling. Given that the dataset has not been applied in this context, it is important to understand the available information. Steps are presented in Figure 10, where the general overview of the dataset and data understanding is described in chapter 3.2, and further results related to data quality are presented in chapter 4.1. Analyzes were performed using Python and Excel, to address data quality, abnormalities, types of data, representativeness, and validation. Data quality is a broad term with many definitions. Chang, Underwood and Roy (2019) refers to data quality as data validity, including its usefulness, accuracy and correctness of data within the intended context. This definition aligns with that of Fleckenstein and Fellows (2018), and thus adopted for this research. To assess data quality, underlying information and variance of the features statistical analysis, knowledge-based review and principal component analysis (PCA) were conducted. PCA involves using a linear combination of the original features, based on maximizing variance (Cohen, 2022). Abdi and Williams (2010) describe PCA as a method to reduce the dimension of data by extracting the most informative features in the dataset, and further analyze the structure in data samples and features.

3.4.3 Data Preparation

The goal of data preparation is to refine the dataset for modelling. A dataset is generally not fully complete and might have information that is not relevant for the exact topic to be researched. This step includes, as seen in Figure 10, taking care of missing values, and selection of the features which contains the relevant information to build a model. Data often requires format conversion to fit the intended model. Guidance on data preparation was taken from “Python Machine learning” by Raschka and Mirjalili (2017) and “Python Data Science Handbook” by VanderPlas (2016). The file format of the dataset was converted from excel to comma-separated-values (.csv) for easier processing of data.

External preprocessing

Rambøll conducted initial preprocessing of the data in advance. As mentioned under section 3.2, there are limited insight in what processes was conducted. Roughly, the data has been adjusted to the right zone with the help of fixed traffic counting spots. This is needed either where telecommunication towers stretch over multiple zones, or with faulty signal traveling. The misclassification happens since the phone signals travel further over water areas which leads them to connect to a tele tower in a different zone than from where it was sent. Further the mobile phone data was merged with the related zone information for the origin-zone (*TotPop*) and destination-zone (*Work, Retail, RecDelPriv*).

Feature selection

The feature selection was conducted in two stages; contextual knowledge and correlation amongst features, as described by Guyon and Elisseeff (2003). Initially, features from the travel survey were removed to focus solely on the value derived from mobile phone data, aligning with the primary business objectives of the research. Duplicate features providing redundant information were also eliminated. The removed features are presented in Table 8.

Table 8 Features removed from the dataset not being evaluated as relevant or sufficient.

Feature	Reasoning
Origin-area	Is the same as origin-zone
Destination-area	Is the same as destination-zone
Ndays	Is similar value for all observations, and only relates to the consistency of the data collection
Walk	Origins from travel survey
Bicycle	Origins from travel survey
Car	Origins from travel survey
Carpass	Origins from travel survey
PT	Origins from travel survey

As suggested by Guyon and Elisseeff (2003), the remaining features were analyzed for correlation. As seen in Figure 11, minimal correlation was observed, except between certain features such as Work and RecDelPriv, and Work and Retail. No additional features were removed based on correlation to preserve information integrity information and be able to evaluate the dataset throughout.

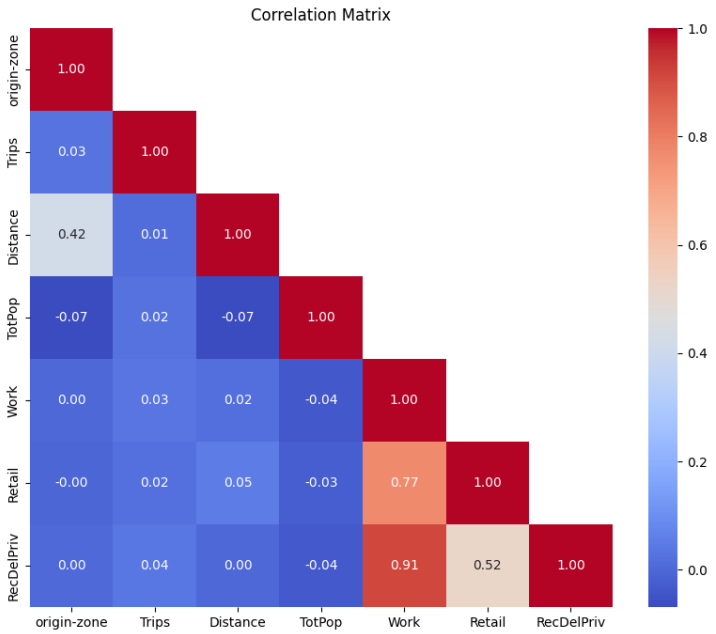


Figure 11 correlation matrix of the features in the dataset.

The remaining selected features is shown in Table 9, where destination-zone is chosen as the response variable. This is in alignment with the overall research question of predicting destination zone, and thereby the algorithms will try to classify the destination based on the information from the other features.

Table 9 The features selected for the data modelling, including the response variable.

Feature	**Comment**
Origin-zone	Origin-zone will give an impact for the
Destination-zone	Response variable for destination choice
Trips	Shows the number of trips between the zone pair
Distance	The travel distance is a known factor for destination choice
TotPop	Known factor for destination choice in transport modelling
Work	Known factor for destination choice in transport modelling
Retail	Known factor for destination choice in transport modelling
RecDelPriv	Known factor for destination choice in transport modelling

Data cleaning

Data cleaning involved removing inconsistencies and converting data types. The features *Trips*, *Distance* and *TotPop* were converted from datatype objects to float numbers, to be applicable for use in a data model. Since the dataset contained no categorical features, no encoding was necessary. Table 10 shows that none of the features in the data set had any duplicates or missing values, showcasing the influence by the external preprocessing conducted on the dataset. The observation gives an indication on high data consistency, leading to a minimal need for data cleaning.

Table 10 Overview of duplicates and missing values for each feature.

	Duplicates	**Missing values (NaN)**
Origin-zone	*None*	*None*
Destination-zone	*None*	*None*
Trips	*None*	*None*
Distance	*None*	*None*
TotPop	*None*	*None*
Work	*None*	*None*
Retail	*None*	*None*
RecDelPriv	*None*	*None*

Inspection of outliers

Outliers can skew a dataset, leading to potential overfitting. On the other side, outliers can contain critical information about anomalies in the dataset. Outliers are identified as values significantly higher or lower than the majority. One of the approaches for outlier detection, suggested by UN council of Experts in Big Data (2023) is The Interquartile Range (IQR) method. Meppelink et al. (2020) has previously used IQR in a mobile phone data context. The calculation for detecting outliers is seen in Equation 3.1 and Equation 3.2. The result from the inspection is shown with violin plots and are presented in chapter 4.1.

$$IQR = Q3 - Q1 \tag{3.1}$$

Where *Q3* is the third quartile percentile and *Q1* is the third quartile.

$$boundary = Q_i + 1.5 * IQR \tag{3.2}$$

Where Q_i is either Q3 for upper boundary or Q2 for lower boundary.

Standardization

In the pipeline all models are run through a standardization with StandardScaler from the Scikit-learn library (Pedregosa *et al.*, 2011), before undergoing hyperparameter tuning with grid search and cross validation, further explained in chapter 3.4.4. Standardization is a statistical process that normalize the feature scales by transforming the feature to an average of zero with a standard deviation of one. This ensures that each feature contributes equally to the analysis (Raschka and Mirjalili, 2017). The formula is given by:

$$Z = \frac{x - \mu}{\sigma} \tag{3.3}$$

Where *Z* is the standardized value, *x* is the dataset value, μ is the average of *x* in the dataset, and σ is the standard deviation of *x* in the dataset (Raschka and Mirjalili, 2017).

3.4.4 Data Modelling

Data modelling involves construction of the model where machine learning models are introduced and trained. The data modelling followed the steps as seen in Figure 10, where models were constructed by doing a selection of models, setting up a pipeline for the model training, conducting a grid search for hyperparameter tuning, then training and testing performance. Each of the models have their unique pipeline with their corresponding grid search parameters. The dataset is divided into an 80% training set and a 20 % testing set**.**

Selection of algorithms

The problem is approached as a multi-class classification problem for destination choice, similar to the methodologies used by Hagenauer and Helbich (2017) and Paredes et al. (2017). The approach was chosen to test how the algorithms were handling the complexity of mobile phone data and focus on prediction accuracy. Alternative methods could be to calculate the probability for choosing each of the destinations, as done by Zhao et al. (2020). The selection of models was based on the presentation in literature, by Raschka and Mirjalili (2017) and VanderPlas (2016), of fundamental different algorithms for supervised learning. For the data modelling a variety of machine learning algorithms presented in Table 11 have been selected; *Logistic Regression*, *Support Vector Machine (SVM), Random Forest* and *Gaussian Naïve Bayes*. The different models were chosen on the basis that they are fundamentally different in their way of training on the data. Where Random Forest is a decision tree algorithm, Logistic Regression is a regression algorithm, SVM is a Support Vector Machine and Gaussian Naïve Bayes is a Naïve Bayes algorithm. The description of the models is presented in chapter 2.4.2.

Table 11 Selected machine learning models

Algorithm	**Scikit-learn name**
Logistic Regression	LogisticRegression
Support Vector Machine	SVC
Random Forest	RandomForestClassifier
Naïve Bayes	GaussianNB

Hyperparameter tuning

Hyperparameter tuning was performed using the GridSearchCV function from Scikit-learn (Pedregosa *et al.*, 2011). Table 12, Table 13, and Table 14 presents the grid search parameters used for hyperparameter tuning. Gaussian Naïve Bayes is not included as it lacks tunable hyperparameters. The strength of model regularization, denoted by Classifier_C, helps to simplify the model. The smaller the C, the stronger the regularization, increasing model complexity at lower values (VanderPlas, 2016). Logistic Regression's solver choice impacts the algorithm used for the loss function, while kernel choice for SVM determines the decision boundary's linearity. Random Forest's max depth and the number of estimators control the complexity and potential overfitting, with a higher number of trees generally enhancing performance but increasing computational demands. The grid search was completed a singular time, whereas the best performing hyperparameters were further applied in a cross-validation.

Table 12 Grid search parameters for Logistic Regression.

Model	C	Solver
Logistic Regression	[0.01, 0.1, 1, 10, 100]	['liblinear', 'lbfgs']

Table 13 Grid search parameters for Support Vector Machine.

Model	C	Kernel
Support Vector Machine (SVM)	[0.1, 1, 10, 100]	['linear', 'rbf']

Table 14 Grid search parameters for Random Froest.

Model	Max depth	N estimators
Random Forest	[None, 10, 20, 30]	[10, 50, 100, 200]

Cross-validation

Cross-validation helps ensure that a model generalizes well across different datasets. It is a method of tackling the bias-variance trade-off. K-fold cross-validation, as illustrated in Figure 12, was applied in this research. In cross-validation one of the folds acts as a validation fold for measuring performance, where the others are used for training. Often leading to a stronger model which is more accurate in the classification, also testing the performance on unseen data (Raschka and Mirjalili, 2017). For the grid-search of each model it was conducted a two-fold cross-validation, and for the second training and testing with optimal hyperparameters it was conducted a five-fold cross-validation. Where the score for the performance metrics was the average score among these five folds.

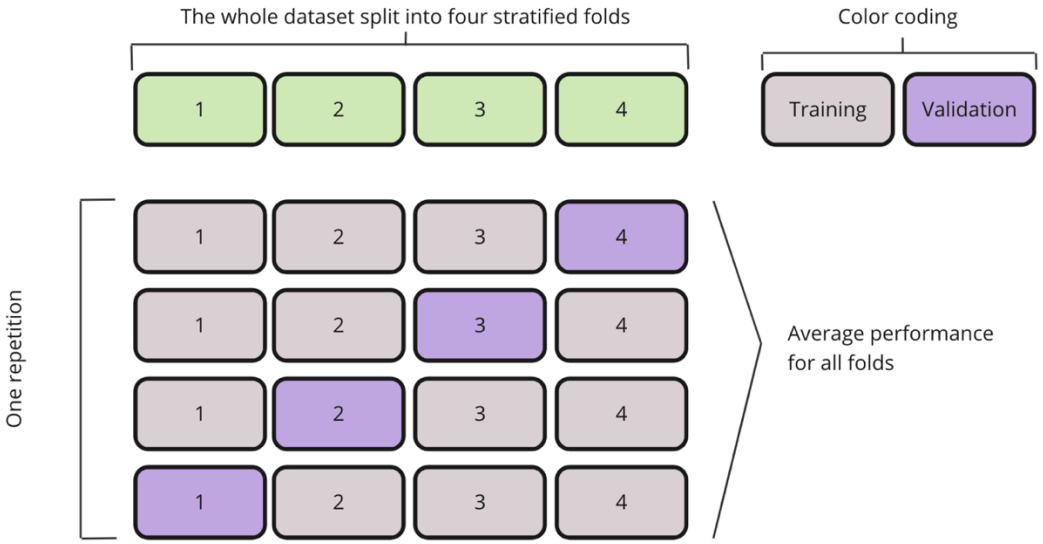


Figure 12: How cross-validation sections a data set into different folds. The result is the average performance for all folds. Inspired by VanderPlas (2016)

Feature importance

Feature importance analysis is conducted for each algorithm to identify which features most significantly impact model predictions. The models have different measures for feature importance, Logistic Regression and SVM measures the coefficient of their decision boundaries, a large value indicates more importance. Random Forest utilize a built-in feature importance function where a large value indicates more importance. Naïve Bayes evaluates the variance of each variable, where a large variance indicates a spread in the feature values that can help the model to make more accurate predictions.

3.4.5 Evaluation

For the evaluation phase models are evaluated based on performance metrics and predetermined criteria. In this research the performance metrics is made for evaluation of the classification models. To get a broader understanding of the performance of the models an addition of qualitative criteria is evaluated through a multi-criteria decision analysis explained in chapter 3.5.

Performance Metrics

Performance metrics quantitatively measure and evaluate the model's relevance. These metrics are derived from the possible outcomes of classifications, which are represented in a confusion matrix in Figure 13. Possible outcomes of a classifying a sample with a classification algorithm are *True Positive* (TP), *True Negative* (FP), *False Positive* (FP) and *False Negative* (FN) (Raschka and Mirjalili, 2017).

		Predicted class: Positive	Predicted class: Negative
True Class	Positive	*True Positive* (TP)	*False Negative* (FN)
	Negative	*False Positive* (FP)	*True Negative* (TN)

Figure 13 Confusion matrix. Inspired by Raschka and Mirjalili (2017)

The selected performance metrics are each calculating the performance on a specific task, giving insight in the balance of the model performance. The metrics chosen are *Accuracy, Precision, Recall, F1-score* and *Matthews Correlation Coefficient (MCC).* Accuracy is chosen for evaluating the prediction accuracy of the models, used in the research of Paredes et al. (2017) and Hagenauer and Helbich (2017) for predictions with travel survey data.

Precision measures how effectively the model predicts positive outcomes. Recall focuses on the model's ability to identify true positives. F1-score is chosen to evaluate the balance between precision and recall in case of skewed results. MCC provides a holistic measure of model performance across all four quadrants of the confusion matrix. Optimal performance is indicated by maximizing these scores.

Accuracy, precision, recall and F1-score gives a score between 0 and 1, and MCC gives a score between -1 and 1. This range of metrics enables a detailed examination of the strengths and weaknesses in the model. To ensure robustness, the mean score across the folds of cross-validation is used, supplemented by the standard deviation of the scores. The calculation of the selected performance metrics is shown in Table 15. These are the quantitative criteria to be used in the MCDA analysis which is further explained in chapter 3.5.

Table 15 Overview of performance metrics formulas and description.

Performance metric	Formula		Description
Accuracy	$ACC = \frac{TP + TN}{FP + FN + TP + TN}$	(3.4)	Accuracy of a model indicates the number of correctly classified samples. It is useful as a general metric for the performance of a model (Raschka and Mirjalili, 2017).
Precision	$PRE = \frac{TP}{FP + TP}$	(3.5)	Precision shows the rate if samples classed as positive that are actually positive (Raschka and Mirjalili, 2017).
Recall	$REC = \frac{TP}{FN + TP}$	(3.6)	Recall is a measure on the fraction of true positive samples in the data set that are classified as true positive (Raschka and Mirjalili, 2017).
F1-score	$F1 = \frac{2 * PRE * REC}{PRE + REC}$	(3.7)	F1-score is a combination of precision and recall used to give an indication of the balance of the two metrics (Raschka and Mirjalili, 2017).
Matthews Correlation Coefficient	$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	(3.8)	MCC combines all four of the classification possibilities for an overall score between -1 and 1. Where 1 indicates everything is correctly classified, -1 is complete opposite classification and 0 is random guess (Chicco and Jurman, 2020).

3.5 Multi-Criteria Decision Analysis

The Multi-Criteria Decision Analysis (MCDA) is used to evaluate the performance of the selection of models tested. Its utility lies in the capability to integrate both quantitative and qualitative criteria, and to assign greater weight to more critical criterions, effectively handling evaluation of alternatives with conflicting objectives (Jordanger *et al.*, 2007). This tool helps identify and give an indication for which models could be further refined and potentially used into transport modelling projects.

3.5.1 Assumptions and Limitations

MCDA inherently involves assumptions and limitations. It is assumed that all models have the possibility to be implemented in the industry, though there might be economic, technological, or other factors that will limit a deployment. When analyzing this research, it is important to take into consideration the assumptions that has been used in the evaluation, as this could be disrupted with the acquisition of new data or new information.

According to the UK Department for Communities and Local Government (2015) it is suggested to limit the complexity of a MCDA with keeping the number of criteria as low as possible, while upholding the ability to take a well-considered decision. In the case-studies by Jordanger et al. (2007) the number of criteria ranges from 5 to 14 for large scale infrastructure projects. For this study, the number of criteria is set to five based on the business objectives understood from the data models. The qualitative criteria were chosen ad-hoc in collaboration with Rambøll, alongside input to which criterion that is considered to be most important. The quantitative criteria were chosen based on input from literature on both transport modelling and data modelling.

The Department for Communities and Local Government (2015) notices that the MCDA is latent to be a subjective model for evaluation. A further risk is that there is no certainty that the correct set of criteria is chosen from the pool of possible criteria, and then also weighted correctly. Therefore, the results from a MCDA have limited use and should be carefully considered and only evaluated in the context where it has been applied. It functions more like a guideline for decision-makers in decisions including interdisciplinary stakeholders, making a more systematic approach to a subjective evaluation.

3.5.2 Evaluation Criteria

The MCDA uses a set of criteria to evaluate the performance of the different models. It is important for the criteria to be grounded in evaluating the performance for meeting objectives (Department for Communities and Local Government, 2015). The chosen set of criteria for this MCDA is *performance, flexibility, simplicity, interpretability,* and *scalability.* Given that deployment of CRISP-DM is not included in the research, the set criteria are more linked to the functions of the models. Still, they are somewhat linked to how the models would function in an operational setting. The criteria are briefly described, further reasoning is provided in Table 16.

Performance assesses the model's ability to accurately predict data classifications, vital for reliable transport planning and decision-making. It is quantitatively measured primarily by accuracy, with additional insights from other performance metrics.

Flexibility measures the model's adaptability to varying data types and conditions without significant modifications. A flexible model can handle variations in the data for both values and types of data.

Simplicity refers to a model's ability to be modified and repaired if faults occur. Simplicity further reduces the risk of overfitting, making the model more generalizable to new data.

Interpretability is important for the model to be used in a decision-making environment. Stakeholders must understand how predictions are made to make an informed decision.

Scalability addresses how well the algorithm can handle increasing amounts of data. With mobile phone data the datasets become large, and for training and testing over multiple geographical areas it can be very computational heavy and should not decrease the performance of the model.

Each criterion is given a weighting based on an evaluation of their importance and criticality for providing industry value within the selected research scope. The sum of the weights must add up to 100%. The weighting is chosen in consideration of literature and discussion with Rambøll, and an the reasoning of the weighting is given in Table 16.

Table 16 Overview of the criteria in the MCDA and the corresponding weights.

Criterion	Weight	Reasoning
Performance	40%	Performance was chosen as the most important criterion as this relates to the main objective of the data models. And is highlighted by Moody (2003) as a good measure of quality of data models and is important for the data analyst in the industry. Also supported Chapman et al. (2000) who says that a model evaluation must be rooted in the business objectives, where in this case accuracy in prediction was a part of both business objectives and research questions. The criterion is further highlighted by Rambøll as important in the weighting process.
Flexibility	10 %	Flexibility is meant by the ability to be adjusted for other data, by adding some more zone information, improve the model further and do adjustment to the model goals. This is a second criterion from Moody (2003), important for the business user in the industry for adjustments of the model goals. Since the industry is applying different data sets and from different areas, not only from mobile phone data, but it will also be of interest to evaluate how flexible the model is. It is not very important since, related to the business objectives, there is no current need for pivoting the model to other types of data or change the model goals.
Simplicity	5 %	Simplicity is much about the data preprocessing required as well as the hyperparameter tuning. In a model it is beneficiary for interdisciplinary teams working with complex problems and is highlighted by Moody (2003) as a measure of quality of data models. To strive for simplicity will make the model easier to work with for further development. It is not weighted very much, due to the willingness of spending time for the processing and tuning of the model for optimal performance.
Interpretability	25 %	The ability to interpret a model is important for the wholistic transport modelling. It is the second most important as the prediction of destination choice is a sub-task in the whole transport planning with the four-step model. Therefore, it is necessary to understand how the model works, under what assumptions and its related decision boundaries. For the transport modelers to gain consistency in the assumptions and quality of work across the whole transport planning. This is a fourth criterion from Moody (2003), important for the business user in the industry. The criterion is further highlighted by Rambøll in the weighting process.
Scalability	20%	Scalability is related to the computational cost and how the literature is evaluating the models to be able to comprehend larger amounts of data. This also relates for the model to be used on data from other geographical areas and ability to integrate into a variety of projects. Rambøll wants this criterion since they are a national and global organization where synergies can be drawn across projects.

3.5.3 Score Giving

Following Jordanger et al. (2007), the scoring scale is set between 1 and 5, suitable for the level of detail achievable in this evaluation. This scale helps to identify broad distinctions among models rather than nuanced differences. For finding the detailed nuances among the alternatives a more detailed description of both criteria and guidance for score giving is necessary, striving to make the qualitative metrics and evaluations anchored in measurable performance. Table 17 shows the guidance used for the score giving for the criteria. The guide was made to give a better insight in what lies behind the evaluation of the different models.

Table 17 Guidance for giving score in the MCDA.

Score	Description of scores	Guidance for score giving
1	Very poor performance	The score 1 is given if the algorithm is not functionable and is not reliable for the given criterion.
2	Poor performance	A poor performance is a below expected performance and experience with the algorithm. It can be scored if the algorithm performs worse than the average of the other algorithms.
3	Moderate performance	If the algorithm has an average performance and there are no distinct advantages experienced in the research or addressed from literature.
4	Good performance	A slightly above average performance given when the algorithm performs well, and literature supports the evaluation of good performance.
5	Very good performance	A top score 5 can be given as a score to the best performer amongst the selection, and thereby the others can be scored relative to that. Since the goal of the MCDA is to distinct the different alternatives among each other, not compared to every existing algorithm. Literature also highlights the algorithm as a top performer for the criterion.

To rank the models a calculation of the total score of the model is done. This is done by multiplying the given score with the weighting.

$$Total\ score = \sum_{i=1}^{n} score_i * weight_i \qquad (3.9)$$

Where, $Total\ score$ is the caculated overall performance of the model based on both quantitative and qualitative criteria. $score_i$ is the score from 1-5 given for the i-th criterion. $weight_i$ is the weighting for the i-th criterion. n is the total number of criteria.

Chapter 4

Results

Chapter 4 presents the results of this research, focusing on the performance on different models. Firstly, it presents results for assessing the data quality of mobile phone data. The chapter concludes with the results from the multi-criteria decision analysis. The aim is to provide a comprehensive overview of the research outcomes before moving to the discussion in chapter 5.

4.1 Quality of Mobile Phone Data

To evaluate the utility of mobile phone data in transport modelling, an analysis of the dataset's features and information was conducted, to further be able to discuss the data quality of mobile phone data.

4.1.1 Statistical Description

As previously noted in Table 10, the dataset demonstrated high consistency with no duplicates or missing values. Table 18 provides a further statistical analysis of each of the features, indicating the distribution of values across the dataset. In most instances, the standard deviation significantly exceeds the mean, where for instance *Trips* has a standard deviation of about ten times the mean. Suggesting a skewed distribution for all features.

Table 18 A statistical overview of the features in the dataset

	Trips	**Distance**	**TotPop**	**Work**	**Retail**	**RecDelPriv**
Mean	487.7	42.6	511.9	276.9	55.1	161.7
Standard deviation	5019.4	43.7	382.8	470.3	129.3	331.8
Min	5.0	0.01	0.0	0.0	0.0	0.0
25 % Quartile	13.8	20.5	210.4	28.0	1.0	6.0
50 % Quartile	50.5	31.5	414.6	103.0	8.0	41.0
75 % Quartile	217.6	45.8	729.6	288.0	42.0	163.0
Max	901785.6	350	2015.0	3603.0	976.0	3557.0

An analysis of zone-pair popularity is presented in Table 19. The results reveal that 63 % of the total 162 517 zone-pairs have 100 or less trips recorded, and less than 0.5 % recorded over 10 000 trips. This suggest that the number of trips per zone-pair could serve as a decisive boundary in the data model for identifying zones with high traffic. However, this approach carries a risk of overfitting due to these outliers.

Table 19 Distribution of zone-pairs across a selection of trip-count intervals

Trips	Number of zone-pairs
>100 000	63
[10 001, 100 000]	660
[1001, 10 000]	12 766
[101, 1000]	47 197
[11, 100]	72 687
[0, 10]	29 144

Table 20 illustrates the number of destination zone pairings per origin zone, indicating the connectivity between zones. The minimum number of connections was two, and zone 480, 431, 438, 481, 467 and 487 all had five or fewer zone pairings. This indicates that the machine learnings models may struggle with these zones due to insufficient data. Conversely, zones 263 and 1325 each had connections to 583 other zones, covering 99% of the zones in the study area, indicating a high level of interconnectivity.

Table 20 Number of destination zone pairings with origin zone.

	Zone pairings
Min	2
Max	584
1st Quartile (25 %)	156
2nd Quartile (50 %)	240
3rd Quartile (75%)	407

4.1.2 Principal Component Analysis

Principal Component Analysis showcase the variance in the features and how they contribute to explanation of linkage between features. Due to computational constraints, the PCA was not performed on the entire dataset. Instead, three distinct analyses were conducted: a loading plot on full dataset, score- and loading plots for all first unique destination zones in the dataset, and score- and loading plot for a sample of the dataset (n = 10 000). Score plots visualize the spread of observations, and the loading plot shows the contribution from each of the different features for the principal components.

The first PCA indicates an explained variance of 54% for principal component 1 (PC1) and 18% for principal component 2 (PC2). The loading plot in Figure 14 shows that *TotPoP* and *Origin-zone* has a substantial influence on PC1, and *Work* and *Origin-zone* significantly impact PC2.

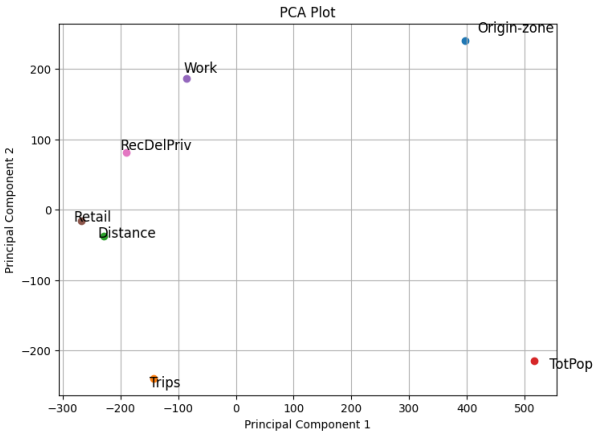


Figure 14 Loading plot of standardized data on all data.

The score plot in Figure 15 has an explained variance by PC1 is 40% and for PC2 is 21%. The loading plot has explained variance of 73% for PC1 and 17% for PC2, somewhat like results in Figure 14. Contrary to Figure 14, the loading plot *Trips* is highly influential on PC2. For the scoring plot it can be detected possible outliers for zone number 481, 480, 63 and 243, indicating regions with distinct data characteristics.

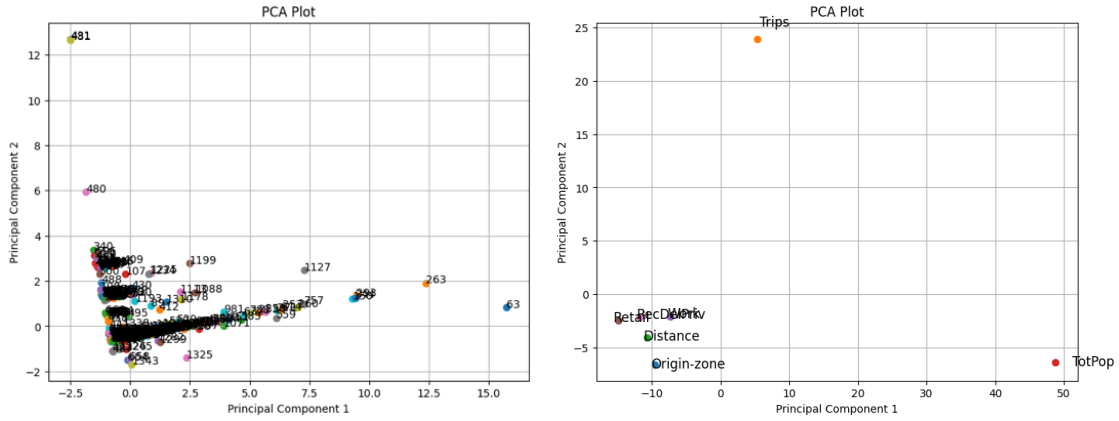


Figure 15 Score- and loading plot of all unique destinations. Score plot to the left.

The score plot in Figure 16 results has explained variance by PC1 is 35% and for PC2 is 21%, a similar result as for Figure 15. Loading plot has explained variance of 54% for PC2 and 18% for PC2, the same results as for the whole dataset shown in Figure 14. In alignment with Figure 15 the loading plot shows that *Trips* is important for PC2, whereas in this case *Origin-zone* is important for PC1 and *TotPoP* is substantially influential on both. The scoring plot detects possible distinguished outlier for zone number 1343 and group of observations related to zone 263.

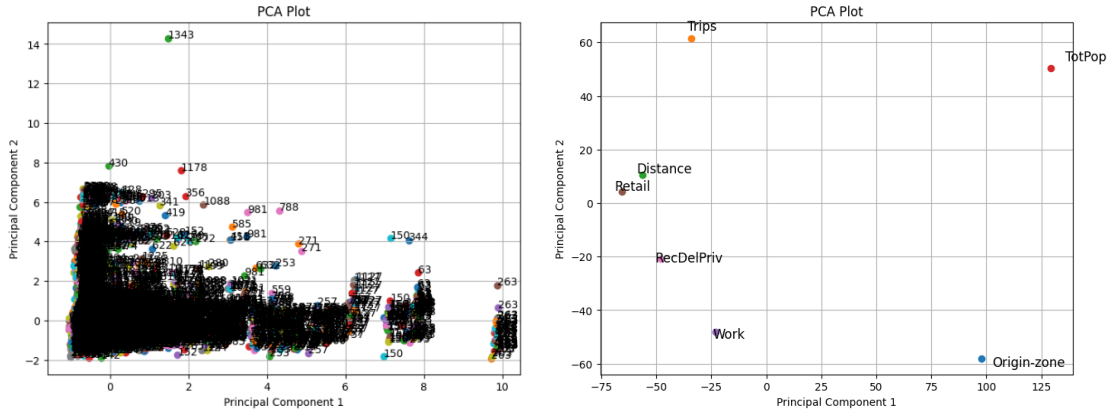


Figure 16 Score- and loading plot for a sample of the data (n = 10 000). Score plot to the left.

The results from each of the principal component analysis show similar levels of explained variance. However, the visual interpretation of data points and feature contributions in the score and loading plots vary. This variation may stem from the different representation of the data within each subsets analyzed.

4.1.3 Violin Plots

Violin plots were employed to analyze the distribution of dataset features in greater detail. The violin plot in Figure 17 illustrates the distribution of the features, except origin-zone and the response variable destination zone. The analysis revealed that most features are highly skewed and contain extreme values. To address this, observations exceeding the upper boundary defined by the Interquartile Range (IQR) formula were removed from each feature, except for TotPop, as shown in Figure 18. Despite modifications, some features – Trips, Work, Retail and RecDelPriv – remained skewed. To further mitigate the risk of overfitting, the IQR adjustment factor was reduced from 1.5*(Q3-Q1) to 0.5*(Q3-Q1) for these features, as seen in Figure 19. The features are left somewhat skewed to avoid excessive removal of potentially important data points and information.

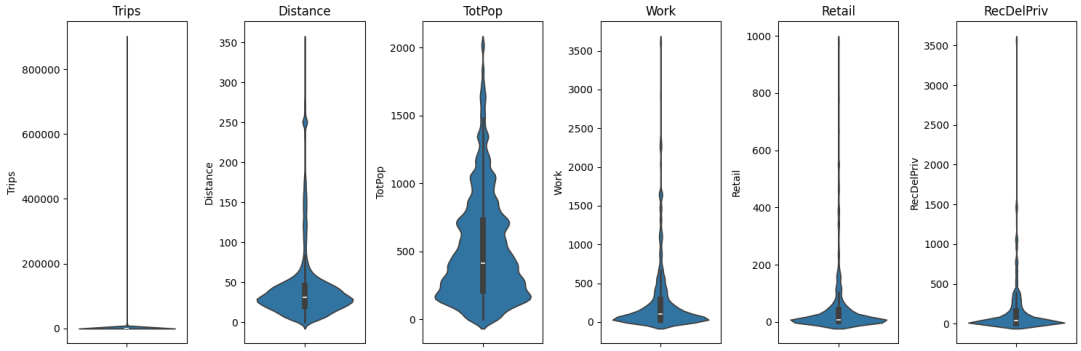


Figure 17 Violin plot of un-treated data.

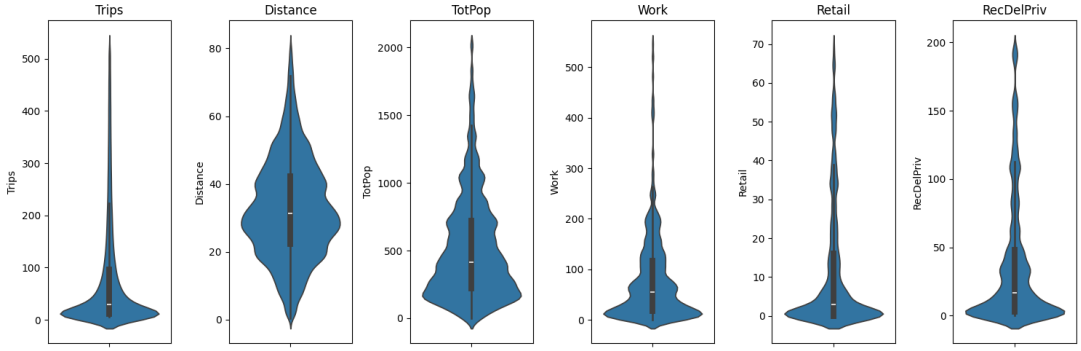


Figure 18 Violin plot, removing extreme values from upper side of features. All except TotPoP.

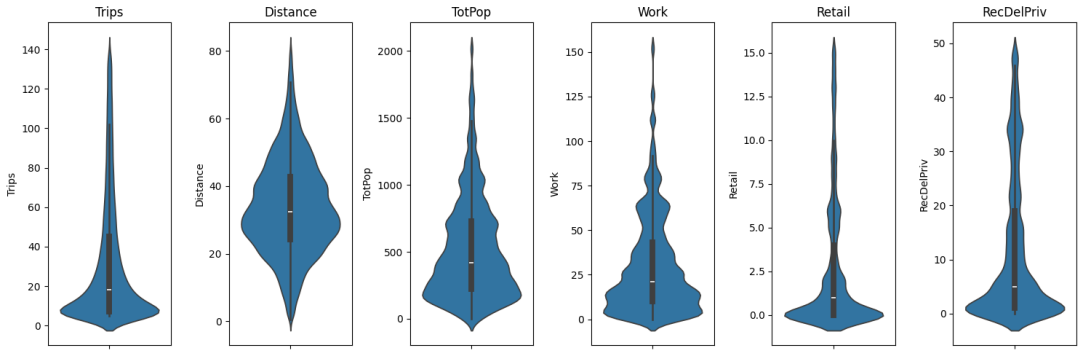


Figure 19 Violin plot removing upper side even more for the features Trips, Work, Retail and RecDelPriv.

4.2 Model Performance

The different models have been constructed and trained on the data before testing up towards the selected performance metrics. This section presents the results from the hyperparameter tuning, analysis on feature importance and quantitative performance.

4.2.1 Best Hyperparameters

Table 21, Table 22, Table 23 and Table 24 present the results from the grid search for hyperparameter tuning, identifying the best hyperparameter combination within the grid-search, and the best accuracy score, which already gives an indication on the best performing model. As indicated in Table 21, Logistic Regression achieved best performance with a high regularization parameter (C), suggesting that the model operates effectively with minimal complexity and that the training data alone sufficiently enables classification. This resembles with Support Vector Machine, which also favored a high C and a linear kernel, showing a preference for simpler models. In contrast, the Random Forest model showed optimal performance with maximal complexity settings in both hyperparameters. Gaussian Naïve Bayes model performed well, suggesting that even a straightforward baseline model is capable of handling the dataset effectively. However, its high performance may also signal potential issues with the dataset or model construction, as Gaussian Naïve Bayes typically shows lower performance compared to more complex models.

Table 21 Overview of Logistic Regression best parameters and cross validation score.

Model	C	solver	Best cross validation score
Logistic Regression	100	lbfgs	0.828

Table 22 Overview of Support Vector Machine best parameters and cross validation score.

Model	C	kernel	Best cross validation score
Support Vector Machine	10	linear	0.914

Table 23 Overview of Random Forest best parameters and cross validation score.

Model	Max depth	N estimators	Best cross validation score
Random Forest	30	200	0.949

Table 24 Overview of Naïve Bayes best parameters and cross validation score.

Model	Best cross validation score
Gaussian Naïve Bayes	0.925

4.2.2 Feature Importance

Table 25 details the feature importance for each model. A higher value indicates greater relevance in the classification of data samples. The results demonstrate that Logistic Regression, Support Vector Machine, and Random Forest primarily value features related to zone data, whereas Naïve Bayes extracts more significant information from features derived from mobile phone data.

Table 25 Shows the feature importance of the features for each tested model.

Feature	Logistic Regression (coefficient)	Support Vector Machine (coefficient)	Random Forest (importance)	Naïve Bayes (Variance)
Work	62.3	0.43	0.322	0.00
RecDelPriv	67.8	0.55	0.246	0.00
Retail	11.3	0.30	0.150	0.00
Distance	0.13	0.00	0.095	121.3
Origin-zone	0.10	0.00	0.082	33959.6
Trips	0.53	0.00	0.059	807.6
TotPop	0.25	0.00	0.046	156373.7

4.2.3 Quantitative Performance

Further analysis of model performance utilized the best-performing hyperparameters in an additional round of cross-validation, with addition of all performance metrics. The results, displayed in Table 26, are score between 0 and 1 (-1 and 1 for MCC) along with the standard deviation across five cross-validation runs. The calculation and description of the metrics is found in chapter 3.4.5. The standard deviation gives an indication on the consistency of the results from the performance metrics. From the results presented in Table 26 (best score is bolded in the table), Random Forest outperforms other models across all metrics. Where Naïve Bayes is coming in second, SVM on third and Logistic Regression as the poorest performer across all metrics.

Table 26 Summary table on the performance of all models with all chosen performance metrics.

Model	Accuracy	Precision	Recall	F1-score	MCC
Random Forest	**0.943 ± 0.001**	**0.885 ± 0.002**	**0.873 ± 0.002**	**0.877 ± 0.002**	**0.943 ± 0.001**
Naïve Bayes	0.925 ± 0.001	0.867 ± 0.000	0.867 ± 0.001	0.860 ± 0.001	0.925 ± 0.001
SVM	0.895 ± 0.001	0.803 ± 0.006	0.798 ± 0.0021	0.791 ± 0.003	0.8943 ± 0.0014
Logistic Regression	0.832 ± 0.002	0.754 ± 0.005	0.722 ± 0.003	0.722 ± 0.005	0.831 ± 0.002

4.3 MCDA Results

This section presents the results from the multi-criteria decision analysis conducted on the algorithms tested in the research. MCDA was selected to get a deeper understanding of the applicability of each model and evaluating the possibility and challenges of mobile phone data in transport modeling. The results are based on a combination of literature review, quantitative performance data, discussions, and empirical insights from the modelling process. An overview of the scoring is provided in Table 27 with the rationale for the score giving presented in the following subsection.

Table 27 The score giving of the MCDA.

Model	Performance (40 %)	Flexibility (10%)	Simplicity (5 %)	Interpretability (25 %)	Scalability (20 %)	Total score
Random Forest	5	5	4	4	4	**4.5**
Naïve Bayes	4	3	5	5	4	**4.2**
Support Vector Machine	4	4	3	3	5	**3.9**
Logistic Regression	3	3	3	5	2	**3.3**

4.3.1 Rationale behind Score Giving

Given that four of the five criteria are qualitative and subjectively assessed, an explanation of the scoring rationale is necessary and presented in this section. The rationale is founded in empirical knowledge from the study and supported by literature.

Performance

Performance is the only quantitative criterion and carries the highest weight. Scores were assigned relative to the top-performing model, which received a score of 5. Models performing at 90-99% of the top model's level were scored 4, and those at 80-89% received a score of score 3. Naïve Bayes being a simple algorithm and VanderPlas (2016) states that it can be used as a baseline algorithm, this indicates that even though the other models are more complex they are not in large degree able to outperform Naïve Bayes. Seen across the result from the other performance metrics the same ranking follows also for them.

Flexibility

Random Forest was rated highest for flexibility due to its ability to adjust to different data types and its range of tunable hyperparameters. The data does not have to be scaled and it can by itself easier utilize the most important features (Raschka and Mirjalili, 2017). Naïve Bayes is flexible in the way that it is simple, however it is not able to tune it toward fitting the data any better. SVM and Logistic Regression need some more tuning than the others. However, SVM is able to use kernel tricks to better handle non-linear classification (VanderPlas, 2016).

Simplicity

In terms of simplicity, Random Forest and Naïve Bayes are both considered user-friendly and nearly "plug and play", with Naïve Bayes being particularly straightforward due to its lack of tunable hyperparameters and its efficiency in training and prediction phases (VanderPlas, 2016). Further Logistic regression is considered a simple algorithm SVM and Logistic Regression, however, require more adjustments and iterations to function effectively, impacting their simplicity scores.

Interpretability

In its simplistic nature Naïve Bayes is very easy to interpret the reasoning behind the classifications (VanderPlas, 2016). Logistic Regression also offers clear interpretability, especially through visualization of its decision-making process. As VanderPlas (2016) and Raschka and Mirjalili (2017) states it can be difficult to interpret the results from a Random Forest if you are looking for the exact reasoning behind the classifications. On the other side, due to the similarity to decision tree, one can easier understand how the model is built up. With SVM it has no direct probabilistic interpretation, which might be difficult to control how the decision boundary is set (VanderPlas, 2016).

Scalability

Concerning scalability, all models faced potential overfitting issues. Logistic Regression encountered additional challenges with computational efficiency and model convergence, earning it the lowest scalability score. Both Random Forest and Naïve Bayes handle large datasets effectively, though Random Forest's performance can degrade as the number of classes increases. SVM, once optimally tuned and trained, offers fast prediction times, indicating good scalability for future applications (VanderPlas, 2016).

The final ranking of the models, presented in Table 28, reflect the aggregated results from the MCDA. Notably, these ranking did not deviate from the initial rankings based solely on performance metrics.

Table 28 Overall ranking of models.

Rank	Model	Total score
1.	Random Forest	4.5
2.	Naïve Bayes	4.2
3.	Support Vector Machine	3.9
4.	Logistic Regression	3.3

Chapter 5

Discussion

In this chapter the research and results from chapter 4 is discussed. It takes into consideration the research questions for the thesis and how the results answer them and builds upon them. Further, it is elaborated on discussions about the chosen methodology. Lastly, presenting topics for further research.

The objective of this study has been to investigate how artificial intelligence in combination with mobile phone data, can improve analysis related to destination choice in transport modelling. Here presenting a reminder of the related research questions, given in chapter 1.3.1:

RQ1: *Are AI-based models capable of utilizing mobile phone data to predict destination choices?*
RQ2: *What opportunities and challenges arise when integrating mobile phone data with artificial intelligence in transport modelling?*

5.1 Mobile Phone Data Prediction Performance

Considering the first research question, the results are strengthening the potential of artificial intelligence combined with mobile phone data in transport modelling. All algorithms tested demonstrated high performance across all quantitative metrics, not only on accuracy. Thill and Wheeler (2000) concludes that a decision tree model is favorable for discrete choice models. As shown Random Forest is also the model to perform best in this research. Which is in alignment with Rahnasto (2022), Hagenauer and Helbich (2017) and Zhao et al. (2010) where the top performer was the Random Forest algorithm, however they used travel survey data. There is scarcity of research on the combination of machine learning and mobile phone data in transport modelling, limiting comparative analysis. However, the good performance of Random Forest can be related to that it is often a high performer in classification algorithms (Prajwala, 2015), and that the function of a Random Forest model closely linked to how a traditional discrete choice model is functioning in transport modelling for destination choice (Sekhar, Minal and Madhu, 2016). Contrary to the findings with travel survey data is that the measured precision is higher on mobile phone data than on some of the research using travel survey data, with a precision of 88.5 % compared to Rahnasto (2022) with 55%. However, the accuracy is close to Hagenauer and Helbich (2017) with a score of just over 90 % for both.

Due to the overall high performance a question that occurs is if the model is able to perform on other geographical areas than Tønsberg. If it is not, the model is not leveraging the strengths and possibilities within machine learning. It limits the model to move towards a purer statistical and mathematical model. Logistic Regression with the simplest hyperparameters and Naïve Bayes were able to perform well, and such simple models should be easy transferable to other areas. The good performance of Naïve Bayes stands in contrast to the findings from Rahnasto (2022), Hagenauer and Helbich (2017) and Zhao et al. (2010). This contrast in performance with previous studies could be attributable to the distinct characteristics of mobile phone data versus travel survey data, warranting further investigation.

The limited insight in how the data was preprocessed before made available in some degree limits the research. An identified preprocessing error is that internal zone travels was labelled with a distance for the zone. The reason was that the internal distance was set to be the minimum distance for traveling from a zone to the nearest zone and back. This was likely done since a trip is not registered until a zone border is crossed, and thus the dwell time led to the trip ending in the origin zone. The discovered faultiness might indicate that there are other faults hidden in the data due to the external preprocessing that is not as easy discoverable.

A promising direction for future research could involve deriving new features from the combination of mobile phone and zone information as a step in the feature extraction of feature engineering (Slimani *et al.*, 2022). Instead of a direct classification it could be better to find the probabilities for travelling to different zones based on origin zone, with further adding more zone information to the origin-zone as well. Deriving new features which linked the zone pairs together. These could for instance be, *TotPop ratio, Work ratio, Retail ratio or RecDelPriv ratio.* Where the ratios give the ratio between the origin-zone and the destination-zone for the given feature. By this it would be possible to find trends in the data that gives an indication for what makes people complete a trip. The trends could investigate travel patterns between low population zones to high population zones or vice versa. By finding these trends it could be easier for it to be generalized across geographical areas. This type of training might be better addressable for learning techniques mentioned in section 2.4.1, about unsupervised learning or reinforcement learning. For a further simplification and generalization of the model the zones could be set into categorical intervals for classifying them as e.g. high-income zones, low population zones and working zones. Then it would be as few as 3-5 different labels, making it much more manageable as a classification problem and less computational heavy.

The machine learning data model will not only increase accuracy of prediction. It would lead to less demand for the costly collection of travel survey data (Anda, Erath and Fourie, 2017). In addition, when a data model is better generalized, it would not be necessary to buy further mobile phone data from teleoperators. The findings of increased accuracy and lower cost for data acquisition resonate with the possible benefits stated by UN council of Experts in Big Data (2023).

As addressed in the results for MCDA, it can be difficult to evaluate different AI models based on qualitative criteria. The method is an experience- and literature-based review of the models, which is not a standardized process for evaluation. It was limited literature comparing algorithms in a transport modelling context, thus decreasing the support in rationale for score giving. Therefore, no hard conclusion can be conducted on the qualitative evaluation of the models. Though, they provide insight in what is to be considered for further development and deployment of models. The evaluation of MCDA as methodology is further discussed in chapter 5.4.1.

5.1.1 Quantitative Performance of Algorithms

The performance across all tested algorithms suggests two possible scenarios: either the dataset features distinct separability, facilitating model accuracy, or there are inherent assumptions or limitations within the model that enhance apparent performance. The data is thought to be separable since every destination zone relates to a unique set of values for *Work, Retail* and *RecDelPriv.* This may allow models to easily identify and utilize patterns for classification.

Another interesting finding is that the algorithms perform well, but utilize different features, as seen in section 4.2.2 on feature importance. For Logistic Regression, Support Vector Machine, and Random Forest they all value the features regarding zone information. Whereas Naïve Bayes is mainly based on the features that origins from mobile phone data. This gives a two-sided indication of both the potential for mobile phone data to be invaluable without merging with zone information, or secondly the possibility to solely rely on mobile phone data.

It would provide extra insight with a simple choice model where each destination choice is solely based on travelling to the nearest zone or a random choice model. Thus, assessing whether complex algorithms significantly outperform simpler heuristic approaches. The performance of even basic models like Naïve Bayes, despite a relatively small number of features, suggest potential unrecognized biases in the dataset or model configurations, contrary to expectations set forth by VanderPlas (2016). Indicating that there are limitations in the model yet to be discovered. Some faulty elements could be choice of wrong response variable, combination of dataset or feature engineering. On the performance side, various literature showcase that the algorithm XGboost in many cases seem to outperform Random Forest which could be of interest to further research in a transport modelling context (Joharestani *et al.*, 2013; Fatima *et al.*, 2023)

5.1.2 The Validity of the Data Model

The primary objective of the data model is to predict the destination zone. Given three of the seven features is derived from the destination zone, they are static for each observation which is linked to that zone. The weakness therefore lies in the transferability to other areas, might indicating less validity of the model. If the model has to be trained for each area the benefits of machine learning diminish. Being limited to the area it has been trained on it would be better to just use a mathematical analysis of the data, looking at the probability for choosing the zones based on the features. Such a mathematical analysis would be a good first iteration towards understanding the value of mobile phone data, not only to serve as a foundational analysis of mobile phone data butt also facilitate a deeper understanding of the data's intrinsic value. However, it could be argued that a machine learning model is just a complex mathematical model that someone else has made for you. Thus, spending time understanding the algorithms will lead to the same understanding of decision-boundaries.

With having a dataset consisting of many different classes the models are very prone to overfit. It seems like the models were able to perform well on the dataset provided, but questions are to be asked on how well it would perform on data from other geographical areas. With the high accuracy on the models, they could most likely benefit from being less restricted. The further hypothesis is that there is more potential in the feature selection and hyperparameter tuning for getting the best trade-off between bias and variance. According to Claesen and De Moor (2015) this can be both a very tedious and costly process but has the potential to improve the performance of the models by a lot.

A further question of the validity is raised within the data processing of mobile phone data and construction of the model. The dataset has in this case been run through multiple instances of processing. For each instance of processing, all the way to the collection of the data, there are made assumptions. With multiple iterations these assumptions might lead to a biased model affected by the data scientist. From the collection there is done assumptions on what zone the signal is related to, due to irregularities in how the signal travels over different landscapes. For instance, Meppelink et al. (2020) has a requirement of 30-minute dwell time, contrary to the 20 minutes Telia operates with. Further they neglect mobile signals from zones larger than 12.5 km in radius, since it was assumed, they contain limited information about the location of a person.

On the other side, Ortuzár and Willumsen (1994) states that assumptions are common in traditional transport modeling, e.g. that all people act rationally and take the route with the lowest generalized cost. UN council of Experts in Big Data (2023) presents biases, precision, and processing requirements as three of the main limitations of mobile phone data. This is due to the nature of the data, where it is high resolution further leads to high noise. Whereas some of the benefits are lower cost for acquiring data compared to travel surveys and the amount of data points. For a more robust approach, engaging directly with the full spectrum of data processing all the way from raw mobile phone data onward, becoming more aware of the validity of the data, which would strengthen the overall validity of this study.

5.1.3 The Quality of Mobile Phone Data

Mobile phone data shows promising results to perform well and be a reliable source for transport modelling. Travel surveys become aggregated data, and Ortuzár and Willumsen (1994) states that all data which is aggregated will have some form of specification error. Mobile phone data, requiring minimal aggregation, potentially reduces these errors, enhancing the precision of transport models.

However, the application of mobile phone data is not without limitations. The necessity to adjust data to fit predefined zones introduces uncertainties, particularly when observations are misassigned and must be corrected using limited static counting stations. This adjustment process can introduce biases associated with the distribution of data across zones. The dataset is limited to a certain geographical area, which may embed local cultural and behavioral biases, potentially limiting the model's applicability to other regions. With the large size of observations, it is not useful to increase the geographical area due to the large, incremental computing cost.

The preprocessing stage of the dataset also revealed challenges; it consisted of very skewed distributions for all of the features. The statistical analysis and Principal Component Analysis showed that there is a potential for outliers in the dataset and that a large degree of the variance could be explained by principal component one and two. It could be of interest to do amore granular examination, solely looking on the raw data to get the full picture of how it is to work with mobile phone data. This is a more incremental step that could lead to a better understanding of the potential and limitations of the dataset, as well as assumptions that are made in the processing of the data. For instance, the research by Montero et al. (2019), used mobile phone data in its raw format to generate origin-destination matrices and verified against known origin-destination matrices.

As the research was addressing the understanding of the data, it became clear that the applications of mobile phone data are very different from other models that has previously been used. It showed that there is a lack of know-how in applying the data. One of the discussed topics was what is wanted to predict, because there was no obvious objective in the data. With destination zone as the response variable it gave over 500 different labels, making it a very complex classification problem. With an increasing number of labels, the model is not as likely to find a generalized pattern for classification (Moral, Nowaczyk and Pashami, 2022). This led to an iterative process for business understanding and data understanding. High level of complexity in the data indicates it beneficial to have a simpler approach for the machine learning.

Operations conducted on the dataset must be carefully considered. The origin-destination pairs are each unique and provide information for finding trends in destination choice. Since the dataset was adjusted for outliers, it implies that unique origin-destination pairs were removed. Especially since the outliers from number of trips were removed it means that the most popular travel routes are excluded from the dataset. Even though this might leave out critical information, it might also be a necessity for not risking overfitting the models. UN council of Experts in Big Data (2023) exemplifies detection of outliers with either Z-score, IQR method or DBSCAN. The choice of IQR resembles with the work of Becker et al. (2013). Other approaches that might give a better detection of outliers since the data was highly skewed is Erman et al. (2007) who used clustering to find abnormal traffic patterns that were significantly different from typical patterns, or Zheng et al. (2008) who used a removing of outliers based on distance, possibly resulting in more nuanced outlier management in skewed datasets.

Ultimately, the integration of AI with mobile phone data in transport modelling presents both opportunities and challenges. This springs out of the complexity of transport planning and transport modelling. It is difficult to solely look at one issue at the same time. For a more detailed insight it would be valuable to separate the trips based on day of week or time of day. Applying this to a model makes it more complex and the insights is not necessarily more valuable, because it leads to many constraints in the model, it is more likely to overfit and not be applicable to new data. The strength of mobile phone data comes in handy if one can simplify and narrow down what to look for in the model, which has been a challenge in the research. This resonates with the topic and research being new to the industry. While mobile phone data offers a potentially lower-cost and more dynamic alternative to traditional data sources, its effective utilization in AI models requires careful consideration of preprocessing, feature engineering, and model training strategies. An iterative approach, possibly starting with the integration of mobile phone data to address specific subtasks within transport modelling, might facilitate a gradual exploration of its value and limitations, and address where traditional datasets and analysis still is necessary.

5.2 Mobile Phone Data Opportunities and Challenges

From a broader perspective, the strength of mobile phone data lies in the high volume of observations, which offers a more representative description of the general population, compared to travel surveys. This large-scale data can significantly enhance the development of accurate origin-destination matrices, serving as a robust statistical foundation for trip generation and further trip distribution in transport models. This capability allows for more precise planning and support in transport modelling contexts.

However, while the quantity of the data is substantial, it's important to address issues like the detail level, computational costs, and interpretability. A weakness related to the accuracy of the count lies in the very short trips that occur within a zone and by definition is not counted as trips. A large amount of data can also still be of poor quality. It is not always that a larger data set will provide better results. As long as the data set is sufficient enough for the algorithms to catch the trends. The challenge is also to correctly put it in a transport model context where it can be a decision support for transport planning. Mobile phone data has for instance limited use for mode choice and detailed route assignments. However, there seems to be sufficient consistency and interpretability of the mobile data to open for extraction of valuable insight for destination choice.

Beyond the scope of destination choice, mobile phone data holds potential for other applications within transport modelling, more elaborated in chapter 5.7 about further research. Mode choice could be relevant given the use of discrete choice models, like destination choice. Transport models able to leverage mobile phone data will have an advantage in their accuracy of models. Gariazzo and Pelliccioni (2019) further utilize the time stamps of mobile phone data to consider daily and seasonal variations in traffic flow. Possibly yielding better prediction of transport demand in high-season and low-season, that could give insight to make solutions for local or interim adjustments for the traffic. A weakness with mobile phone data is the lack of providing characteristics of the person travelling. A travel survey adds features like age, sex and occupation. This data is often needed in other transport analysis and for the overall transport model estimation.

In the context of integrating artificial intelligence with mobile phone data, it is essential to clearly define the business objectives before developing the data model. Deciding on the business objectives would make it easier to construct a data model that solves the objective. It can be argued that artificial intelligence could be replaced by simple mathematical probability calculations for the origin-destination pairs, more like Essadeq and Janik (2021). However, artificial intelligence is in fact mathematical operations done at a higher and more complex level. Therefore, a mathematical probability model could be a precursor to an AI-model as the first iteration for making a sufficient destination choice model with high degree of interpretability.

5.3 Mobile Phone Data Separation from Travel Survey

The research suggests that reliance on travel surveys for destination choice modeling may not be as necessary as previously thought. While survey provides a deeper insight and broader information, in terms of characteristics of the person travelling, solely looking at destination choice could indicate relying only on mobile phone data will be sufficient. This is contrary to Caceres, Romero and Benitez (2020), which is more focused on the data fusion where a travel survey can equalize for the weaknesses of mobile phone data.

As a holistic approach to transport modelling, it could be eligible to solely use a travel survey. Since this gives insight in trip generation, trip distribution and mode choice. Therefore, the cost-benefit for making the additional work of buying and aggregating the mobile data must be evaluated. A weighting of how a survey or mobile phone data has the ability to generalize, and scale should also be included in the cost-benefit analysis. As referred to, the work by Hagenauer and Helbich (2017) and Zhao et al. (2010) shows that travel survey data is also well compatible with machine learning for destination predictions.

Seen in light of Equation 2.1, for the generalized cost, it shows it to be reliant on multiple data sources, one of them being travel surveys, for estimating the coefficients of each variable. Before again estimating the probability of destination choice based on the generalized cost for the person travelling. In contrast, mobile phone data allows for direct observation-based modelling, eliminating the need for some traditional estimation steps. This direct approach could streamline the modelling process but may require sophisticated analytical techniques to effectively extract meaningful insights without the contextual depth provided by travel surveys.

5.4 Choice of Methodology

The CRISP-DM framework, adopted for this research, demonstrates advantages and limitations. It provides a structured approach to data mining, emphasizing the importance of aligning the analysis with business objectives. An often occurring problem is data mining becoming too focused on the data and data model, forgetting about the business-related goals. A challenge with CRISP-DM is its descriptive nature about what should be done without prescribing how to execute these steps. This opens for an interpretation of what the steps include and might not lead to a good enough continuity among different research that is conducted with this methodology. On the other side, CRISP-DM, with its industry near approach, limits this gap between research and operational implementation, possibly shortening the time for a new technology to benefit society. This sprung from the emphasis the framework puts on the iterative process between data understanding and business understanding. This study has shown that when applying a new type of data set into an industry it is important to iterate on how a model can be designed with the data, and much more important how the model is providing value for transport planning.

If the research were to apply a different methodology it could have given some different results. CRISP-ML(Q) could possibly have given a more pointed approach for modelling with machine learning. This approach could better address how the data could be processed more tailored for machine learning. However, with CRISP-DM being a paramount framework it could be argued for the more nuanced frameworks to be merged with CRISP-DM steps. Biz-ML by Siegel (2024) is interesting for a more deployment focus of the research, but too preliminary to be set in a research context. Given that the focus on the deployment was excluded from CRISP-DM, the biz-ML would not provide additional value for the objectives of this research. Other studies are quite unanimous and agrees that big data and artificial intelligence is the future for transport modelling. Therefore, it would be of interest for the further research to scope closer to deployment.

Different results could occur by constructing the model in a standardized academic approach, contrary to industry near. Possibly leading to better understanding of the steps conducted in the methodology and have a more robust documentation. On the other side, it would risk of narrowing down the scope too much and solely looking at the algorithms and dataset without considering business objectives and a holistic evaluation of models. The conducted grid-search also showcase limitations for the models to tune for best performance, especially since two of the algorithms used max values for one or more hyperparameters. Thus the grid-search should be extended.

With the MCDA as the chosen method for evaluating the model it gave further dimensions to the framework. One could also argue that using it in a case study, prototype or for an actual project would give very valuable insight for the score giving of the criteria from the MCDA.

5.4.1 MCDA as tool for evaluation of algorithms

Using MCDA as the principal method for evaluating and ranking the models significantly impacts the research outcomes. The construction of the analysis with choice of criteria and weighting as the most influential variables, yet they introduce subjective elements that could affect the robustness of the findings. There is a subjective choice of criteria and a subjective choice of weighting of the different criteria, which cannot be guaranteed to be optimal. Conducting a deployment test of the models could validate the chosen criteria and weightings, potentially leading to adjustments based on real-world applicability.

A robust enhancement for the MCDA could include conducting a sensitivity analysis through a Monte-Carlo simulation, as done by Ambrasaite, Barfod and Salling (2011). This could be conducted on both weights and score giving. This approach would help in solidifying the MCDA's credibility, especially when dealing with a novel data type like mobile phone data. Since working with mobile phone data is quite new it was beneficial to add some qualitative criteria and evaluate the performance on the models both quantitatively and qualitatively. By incorporating this it was easier to evaluate how it was to work with that type of data, as it's seen in an operational context, as well taking into consideration conflicting objectives (Montis *et al.*, 2000). Even though the criterion with the highest weighting were solely quantitative, one could strive to find quantitative metrics on the other criteria as well, such as number of lines of code and computational time. This would leave to a less biased evaluation of the models. A further bi-effect of the MCDA that Montis et al. (2000) mentions is the increase of participation and understanding of the model by stakeholders and decision-makers.

The MCDA gave extra insight in how the algorithms are to work with, which is important for a potential implementation. In some terms one can say it is linked to the quality assurance of the CRISP-ML(Q) and therefore closing the gap between the nuances between the two frameworks. The Department for Communities and Local Government (2015) states that a key feature of the MCDA is the judgement of the decision-making team in the establishment of objectives, criteria, and weighting.

In conclusion, while the MCDA offers valuable insights to the practicalities of working with different algorithms, its subjective nature opens the need for careful consideration in a standardized research setting.

5.5 Research Limitations and Limitations of Findings

The findings from this research primarily identifies which machine learning models appear most effective using mobile phone data. However, the algorithms are not thoroughly tested for optimal performance. Further analysis on hyperparameter tuning and feature engineering might unlock more potential in the mobile phone data. Additionally, the data is limited to findings from Tønsberg, Norway, means the findings may not directly apply to other regions where local, cultural differences may occur. These patterns are thought to be because of socio-economic differences, mode choice, architecture of established infrastructure and city design.

The MCDA utilized in this research serves more as a guideline for future modelling priorities rather than providing a definitive evaluation of the models. This approach highlights the need for cautious interpretation of the MCDA results and suggest that future research should include more robust validation methods to confirm the findings.

5.6 Contribution: Implications in Transport Modelling

For industry professionals, this research advances the integration of artificial intelligence models and mobile phone data into transport modelling projects. The results reinforce the hypothesis that AI-models are beneficial in this context and advocate for continued research focused on the operational and deployment phases of model development.

It is important to notice that mobile phone data needs to be managed correctly and will therefore increase the demand for data science knowledge for engineers in transport modelling. The models will become more complex since a larger amount of data is applied, as well as a relatively new data source. For them all to function together it will be managerial challenges. These challenges require skilled leadership capable of guiding interdisciplinary teams and navigating the assumptions underlying in the data processing and model construction.

While this study does not delve deeply into compliance issues, it is critical to also acknowledge that the use of mobile phone data in transport modelling must adhere to GDPR regulations. From a socio-economic perspective the implementation of better predictive analysis of destination choice will lead to cost savings, not only for the time spent for engineering the solution, but also the long-term savings on constructing and building better solutions for infrastructure that more accurately meets societal needs.

5.7 Further Research

The study has provided valuable insights in the research directed towards application of artificial intelligence in transport modelling. Being a novel iteration of the application of mobile data there are numerous challenges to address and opportunities to explore. The suggested further research is split into three topics where one is focused on the mobile phone data itself, the second is focused on other artificial intelligence models that can provide better results, and the third is focused on the deployment of AI/ML-models in transport modelling.

Maximizing utilization of mobile phone data

Future research should investigate ways to standardize the collection and processing of mobile phone data with fewer assumptions and reduced reliance on other data sources for validation. An in-depth analysis of trends in origin-destination zone pairs could reveal common indicators that influence travel behaviors, investigating what makes a person travel from one zone to another. Utilizing raw mobile phone data integrated with zone information might allow for the development of probabilistic models rather than strict classification models. Such models, could potentially excel with a refined approach, offering broader applicability across different geographic regions and yielding insights that could inform general transport modelling strategies This could be accomplished by merging mobile phone data with the zone information for both origin and destination zone. Then all observations would be a singular trip, and the model could easier separate the data for extracting trends. By this it is believed that the model would be more transferable to different geographical areas and gain valuable insight for overall transport modelling. With the further possibility of deriving new features from the data.

Other artificial intelligence models

There is significant scope to expand the range of artificial intelligence models tested on mobile phone data beyond those evaluated in this study. Future studies could explore unsupervised learning techniques or reinforcement learning to enhance data interpretability and discover underlying patterns without predefined labels. Neural networks can perform well on mobile phone data due to the high number of observations and labels related to the dataset. It would also be beneficial to benchmark these sophisticated models against simpler algorithms to identify any hidden biases or data issues that might not be apparent from high-performing models alone.

Machine learning models ready for deployment

While the potential of integrating mobile phone data and machine learning in transport modelling is recognized, few other studies, including this, have any concrete points to how the models can be implemented into an operational role. Further research should apply CRISP-DM, or CRISP-ML(Q) with a focus on the sixth step, deployment of the model. Assessing barriers of entry and necessary steps for the industry to apply it in projects. Further, it could be investigated how machine learning models can be integrated into existing transport modelling software. Additionally, conducting cost-benefit analyses comparing travel surveys and mobile phone data could address the most economically viable data sources for transport modelling. Lastly, implementing pilot projects would provide empirical insights into the challenges and requirements for embedding these models into a project.

These research directions not only aim to improve the theoretical understanding and application of Ai in transport modelling but also seek to bridge the gap between academic research and practical, operational use, ultimately enhancing the efficiency and quality of transport infrastructure planning and development.

Chapter 6

Concluding Remarks

This study aimed to investigate the potential of artificial intelligence-based models using mobile phone data from Tønsberg for transport modelling. The research scope was to particularly focus on the accuracy of destination choice prediction, and to evaluate the associated challenges and opportunities of leveraging mobile phone data for artificial intelligence in this field. The study provided a comprehensive theoretical framework, detailed the research process, described the development of the data, before presenting and discussing the acquired results.

Firstly, it was looked at the quality of mobile phone data in a machine learning context. The research supports other literature that showcase the positive potential for utilizing mobile phone data. However, challenges remain in ensuring the interpretability of models and maintaining clarity regarding the assumptions embedded within the data and modelling process.

Secondly, a selection of machine learning algorithms was evaluated for their ability to predict destination choices using mobile phone data. The study supports other literature on the increased prediction accuracy when applying machine learning algorithms. With Random Forest coming out as the top performer with an accuracy of 0.943. Surprisingly, Naïve Bayes coming in at second with an accuracy of 0.925, contrary to other literature. Followed by Support Vector Machine with 0.895 and lastly Logistic Regression with 0.832. However, all models had quite similar performance, all on a highly acceptable level.

Thirdly, the algorithms were further evaluated with a multi-criteria decision analysis. The conducted MCDA provided a deeper understanding of each model's potential for application in transport modelling. This evaluation method confirmed the initial performance rankings and revealed distinct strengths and weaknesses of the algorithms. It is highlighted that the MCDA could benefit from increased robustness of evaluation.

Further investigations should be considered about the validity of the data model, as the performance is abnormally high across all models. Therefore, further research is needed to both validate the results, examine the model's transferability to other geographical areas and explore practical implementation strategies within the infrastructure sector. This research contributes to the evolving field of transport modelling by integrating advanced AI techniques with emerging data sources like mobile phone data, paving the way for more sophisticated, data-driven approaches in infrastructure development.

References

Abdi, H. and Williams, L.J. (2010) 'Principal component analysis', *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4). Available at: https://doi.org/10.1002/wics.101.

Ambrasaite, I., Barfod, M.B. and Salling, K.B. (2011) 'MCDA and risk analysis in transport infrastructure appraisals: The Rail Baltica case', *Procedia - Social and Behavioral Sciences*, 20. Available at: https://doi.org/10.1016/j.sbspro.2011.08.103.

Anda, C., Erath, A. and Fourie, P.J. (2017) 'Transport modelling in the age of big data', *International Journal of Urban Sciences*, 21. Available at: https://doi.org/10.1080/12265934.2017.1281150.

Andersson, A., Winslott Hiselius, L. and Adell, E. (2018) 'Promoting sustainable travel behaviour through the use of smartphone applications: A review and development of a conceptual model', *Travel Behaviour and Society* [Preprint]. Available at: https://doi.org/10.1016/j.tbs.2017.12.008.

Azevedo, A. and Santos, M.F. (2008) 'KDD, SEMMA and CRISP-DM: A parallel overview', *IADIS European Conference Data Mining 2008*, (January 2008), pp. 182–185. Available at: https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview (Accessed: 10 April 2024).

Becker, R. *et al.* (2013) 'Human mobility characterization from cellular network data', *Communications of the ACM*, 56(1), pp. 74–82. Available at: https://doi.org/10.1145/2398356.2398375.

Bloch, V.V.H. (2024) *Standard for delområde- og grunnkretsinndeling*. Available at: https://www.ssb.no/klass/klassifikasjoner/1 (Accessed: 6 May 2024).

Caceres, N., Romero, L.M. and Benitez, F.G. (2020) 'Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys', *Transportmetrica A: Transport Science*, 16(3), pp. 574–601. Available at: https://doi.org/10.1080/23249935.2020.1720857.

Cargill, M. and O'Connor, P. (2009) *Writing Scientific Research Articles: Strategy and Steps*, *Veterinary Pathology*. West Sussex: Wiley-Blackwell. Available at: https://doi.org/10.1177/0300985813501283.

Chang, W., Underwood, M. and Roy, A. (2019) 'NIST Big Data Interoperability Framework: Volume 4, Security and Privacy', *NIST Special Publication*, 4, pp. 1–176. Available at: https://doi.org/https://doi.org/10.6028/NIST.SP.1500-4r2.

Chapman, P. *et al.* (2000) 'Step-by-step data mining guide', *SPSS inc*, 78, pp. 1–78. Available at: http://www.crisp-dm.org/CRISPWP-0800.pdf (Accessed: 18 March 2024).

Chicco, D. and Jurman, G. (2020) 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21(1), pp. 1–13. Available at: https://doi.org/10.1186/s12864-019-6413-7.

Claesen, M. and De Moor, B. (2015) 'Hyperparameter Search in Machine Learning', pp. 10–14. Available at: http://arxiv.org/abs/1502.02127 (Accessed: 7 May 2024).

Cohen, M.X. (2022) *Practical Linear Algebra for Data Science From Core Concepts to Applications Using Python*. Sebastopol: O'Reilly Media Inc.

Corrales, D.C., Ledezma, A. and Corrales, J.C. (2015) 'A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal', *Journal of Computers*, 10(6), pp. 396–405. Available at: https://doi.org/10.17706/jcp.10.6.396-405.

Department for Communities and Local Government (2015) *Multicriteria Analysis - A manual*. Available at: http://eprints.lse.ac.uk/12761/1/Multi-criteria_Analysis.pdf (Accessed: 10 April 2024).

Erman, J. *et al.* (2007) 'Identifying and discriminating between web and peer-to-peer traffic in the network core', *16th International World Wide Web Conference, WWW2007*, pp. 883–892. Available at: https://doi.org/10.1145/1242572.1242692.

Essadeq, I. and Janik, T. (2021) 'Use of Mobile Telecommunication Data in Transport Modelling - A French Case Study', *International Transport Forum* [Preprint]. Available at: www.itf-oecd.org (Accessed: 23 April 2024).

European Commission (2008) *Guide to cost benefit analysis of investment projects*. Available at: https://doi.org/10.11646/zootaxa.1910.1.6.

Fatima, S. *et al.* (2023) 'XGBoost and Random Forest Algorithms: An in Depth Analysis', *Pakistan Journal of Scientific Research*, 3(1), pp. 26–31. Available at: https://doi.org/10.57041/pjosr.v3i1.946.

Fleckenstein, M. and Fellows, L. (2018) *Modern Data Strategy*. Cham: Springer International Publishing.

Fortmann-Roe, S. (2012) *Understanding the bias-variance trade-off*. Available at: https://scott.fortmann-roe.com/docs/BiasVariance.html (Accessed: 15 April 2024).

Frazier, P.I. (2018) ‘A Tutorial on Bayesian Optimization’, (Section 5), pp. 1–22. Available at: http://arxiv.org/abs/1807.02811 (Accessed: 1 May 2024).

Gariazzo, C. and Pelliccioni, A. (2019) ‘A Multi-City Urban Population Mobility Study Using Mobile Phone Traffic Data’, *Applied Spatial Analysis and Policy*, 12(4), pp. 753–771. Available at: https://doi.org/10.1007/s12061-018-9268-4.

Gilbert, J.K. (2004) ‘Models and modelling: Routes to more authentic science education’, *International Journal of Science and Mathematics Education*, 2(2). Available at: https://doi.org/10.1007/s10763-004-3186-4.

Google (no date) *Google Colaboratory*. Available at: https://colab.google/ (Accessed: 29 April 2024).

Goves, C. and Hemmings, T. (2016) ‘Utilising mobile phone data for transport modelling’, pp. 1–74. Available at: www.ts.catapult.org.uk (Accessed: 6 May 2024).

Guyon, I. and Elisseeff, A. (2003) ‘An Introduction to Variable and Feature Selection’, *Journal of Machine Learning Research 3* [Preprint]. Available at: https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf (Accessed: 10 May 2024).

Hagenauer, J. and Helbich, M. (2017) ‘A comparative study of machine learning classifiers for modeling travel mode choice’, *Expert Systems with Applications*, 78, pp. 273–282. Available at: https://doi.org/10.1016/j.eswa.2017.01.057.

Harris, C.R. *et al.* (2020) ‘Array programming with NumPy.’, *Nature*, 585(7825), pp. 357–362. Available at: https://doi.org/10.1038/s41586-020-2649-2.

Hoff, I. and Nordahl, R.S. (2024) *Vei*, *Store norske leksikon*. Available at: https://snl.no/vei (Accessed: 18 March 2024).

Hunter, J.D.; (2007) ‘Matplotlib: A 2D Graphics Environment’, *Computing in Science & Engineering*, 9(3), pp. 90–95. Available at: https://doi.org/http://dx.doi.org/10.1109/MCSE.2007.55.

Indahl, U.G. and Mintorovitch, N.P. (2024) *Guidelines for use of artificial intelligence at REALTEK*. Available at: https://www.nmbu.no/en/faculties/faculty-science-and-technology/kunstig-intelligens-ved-realtek (Accessed: 8 April 2024).

Johannessen, A., Christoffersen, L. and Tufte, P.A. (2020) *Forskningsmetode for økonomisk-administrative fag*. 4th editon. Oslo: Abstrakt forlag.

Joharestani, M.Z. *et al.* (2013) 'Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data', (1992), pp. 6425–6432. Available at: https://doi.org/https://doi.org/10.3390/atmos10070373.

Jordanger, I. *et al.* (2007) *Concept rapport nr. 18: Flermålsanalyser i store statlige investeringsprosjekt*. Available at: http://hdl.handle.net/11250/228091 (Accessed: 26 April 2024).

Jupyter (no date) *Jupyter*. Available at: https://jupyter.org/ (Accessed: 29 April 2024).

MathWorks (2024) *Overfitting*. Available at: https://www.mathworks.com/discovery/overfitting.html (Accessed: 1 May 2024).

MathWorks (no date) *What Is a Support Vector Machine?* Available at: https://www.mathworks.com/discovery/support-vector-machine.html (Accessed: 24 April 2024).

McKinney, W. (2010) 'Data Structures for Statistical Computing in Python', *Proceedings of the 9th Python in Science Conference*, 1(Scipy), pp. 56–61. Available at: https://doi.org/10.25080/majora-92bf1922-00a.

Meppelink, J. *et al.* (2020) 'Beware thy bias: Scaling mobile phone data to measure traffic intensities', *Sustainability (Switzerland)*, 12(9). Available at: https://doi.org/10.3390/su12093631.

Ministry of Justice and Public security (2018) *Act relating to the processing of personal data (LOV-2018-06-15-38)*, *Lovdata*. Available at: https://lovdata.no/dokument/NL/lov/2018-06-15-38 (Accessed: 30 April 2024).

Montero, L. *et al.* (2019) 'Fusing mobile phone data with other data sources to generate input OD matrices for transport models', *Transportation Research Procedia*, 37(September 2018), pp. 417–424. Available at: https://doi.org/10.1016/j.trpro.2018.12.211.

Montis, A. De *et al.* (2000) 'Criteria for quality assessment of MCDA methods', *3rd Biennial Conference of the European Society for Ecological Economics,* (January), p. 30. Available at: https://www.researchgate.net/publication/228729314_Criteria_for_quality_assessment_of_MCDA_methods (Accessed: 10 April 2024).

Moody, D.D.L. (2003) 'Measuring the Quality of Data Models: An Empirical Evaluation of the Use of Quality Metrics in Practice', *ECIS 2003 Proceedings.*, p. Paper 78. Available at: http://is2.lse.ac.uk/asp/aspecis/20030099.pdf (Accessed: 29 April 2024).

Moral, P. Del, Nowaczyk, S. and Pashami, S. (2022) 'Why Is Multiclass Classification Hard?', *IEEE Access*, 10, pp. 80448–80462. Available at: https://doi.org/10.1109/ACCESS.2022.3192514.

NENT - The National Committee for Research Ethics in Science and Technology (2019) 'Statement on Research Ethics in artificial intelligence'. Available at: https://cha-shc.ca/english/about-the-cha/statement-on-research-ethics.html (Accessed: 30 April 2024).

Nilsen, Ø.L., Uteng, A. and Myrberg, G. (2021) *Mobilitetskartlegging Tønsberg*. Available at: https://www.tonsberg.kommune.no/_f/p1/i83e37574-5ed0-4d30-b5f2-e9529c2a6312/kartlegging-av-mobilitet-i-ny-kommune.pdf (Accessed: 30 April 2024).

Norwegian Ministry of Transport (2024) *Meld. St. 14 (2023 – 2024) - Nasjonal Transportplan 2025-2036*. Available at: https://www.regjeringen.no/no/dokumenter/meld.-st.-14-20232024/id3030714/ (Accessed: 3 April 2024).

Norwegian Road Law (LOV-1963-06-21-23) (1963) *Lovdata*. Available at: https://lovdata.no/dokument/NL/lov/1963-06-21-23 (Accessed: 14 April 2024).

Odeck, J. and Welde, M. (2015) 'Resource allocation in the transport sector : some potential improvements', *Concept report*, (44), p. 17. Available at: https://www.ntnu.no/documents/1261860271/1262010703/Concept_No44_eng_ny.pdf/c8f2cb3c-2c92-4214-ae70-c3b260282327 (Accessed: 21 April 2024).

OpenAI (2024a) *ChatGPT*. Available at: https://chat.openai.com/ (Accessed: 30 April 2024).

OpenAI (2024b) *ChatGPT Overview*. Available at: https://openai.com/chatgpt (Accessed: 30 April 2024).

Ortuzár, J. de D. and Willumsen, L.G. (1994) *Modelling Transport*. second edi. West Sussex: John Wiley & Sons Ltd.

Ortúzar, J. de D. and Willumsen, L.G. (2011) *Modelling Transport*. second edi, *Modelling Transport*. second edi. West Sussex: John Wiley & Sons Ltd. Available at: https://doi.org/10.1002/9781119993308.

Owen, L. (2022) *Hyperparameter Tuning with Python: Boost your machine learning model's performance via hyperparameter tuning*. Birmingham: Packt Publishing Ltd.

Oxford English Dictionary (2002) *Model*. Available at: https://www.oed.com/dictionary/model_n?tl=true (Accessed: 1 May 2024).

Paredes, M. *et al.* (2017) 'Machine learning or discrete choice models for car ownership demand estimation and prediction?', *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017 - Proceedings*, pp. 780–785. Available at: https://doi.org/10.1109/MTITS.2017.8005618.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, 12(May 2014), pp. 2825–2830. Available at: https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python (Accessed: 29 April 2024).

Prajwala, T.R. (2015) 'A Comparative Study on Decision Tree and Random Forest Using R Tool', *Ijarcce* [Preprint], (January 2015). Available at: https://doi.org/10.17148/ijarcce.2015.4142.

Python (2023) *Python 3.9.18*. Available at: https://www.python.org/downloads/release/python-3918/ (Accessed: 29 April 2024).

Rahnasto, I. (2022) 'Comparing discrete choice and machine learning models in predicting destination choice'. Available at: https://urn.fi/URN:NBN:fi:aalto-202206194015 (Accessed: 14 April 2024).

Raschka, S. and Mirjalili, V. (2017) *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn and TensorFlow*. Birmingham: Packt Publishing Ltd.

Rolstadås, A. *et al.* (2021) *Praktisk prosjektledelse*. 2nd edn. Bergen: Fagbokforlaget.

Sekhar, Ch.R., Minal, M. and Madhu, E. (2016) 'Multimodal Choice Modeling Using Random Forest Decision Trees', *International Journal for Traffic and Transport Engineering*, 6(3), pp. 356–367. Available at: https://doi.org/10.7708/ijtte.2016.6(3).10.

Shoman, W. *et al.* (2023) 'A Review of Big Data in Road Freight Transport Modeling: Gaps and Potentials', *Data Science for Transportation*, 5(1), pp. 1–16. Available at: https://doi.org/10.1007/s42421-023-00065-y.

Siegel, E. (2024) *Getting Machine Learning Projects from Idea to Execution*. Available at: https://hbr.org/2024/01/getting-machine-learning-projects-from-idea-to-execution (Accessed: 10 April 2024).

Slimani, I. *et al.* (2022) 'Automated machine learning: the new data science challenge', *International Journal of Electrical and Computer Engineering*, 12(4), pp. 4243–4252. Available at: https://doi.org/10.11591/ijece.v12i4.pp4243-4252.

Store Norske leksikon (2024) *Tønsberg*. Available at: https://snl.no/Tønsberg (Accessed: 9 May 2024).

Studer, S. *et al.* (2021) 'Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology', *Machine Learning and Knowledge Extraction*, 3(2), pp. 392–413. Available at: https://doi.org/10.3390/make3020020.

Svaboe, G.B.A. (2024) 'Travel survey methodology: Advantages, disadvantages, and unintended side-effects of survey design choices'. Available at: https://hdl.handle.net/11250/3111590 (Accessed: 22 April 2024).

Tefficient (2022) *Assessment of Norwegian mobile revenues in a Nordic context – 2022*. Available at: https://www.regjeringen.no/no/dokumenter/assessment-of-norwegian-mobile-revenues-in-a-nordic-context-2022/id2909625/ (Accessed: 22 April 2024).

Telia (2024a) *Crowd Insight for transport*. Available at: https://www.telia.no/bedrift/crowd-insights/crowd-insights-for-transport/ (Accessed: 10 May 2024).

Telia (2024b) *Personvern først. Alltid.* Available at: https://www.telia.no/bedrift/crowd-insights/personvern-forst-alltid/ (Accessed: 10 May 2024).

Thill, J.C. and Wheeler, A. (2000) 'Tree induction of spatial choice behavior', *Transportation Research Record*, (1719), pp. 250–258. Available at: https://doi.org/10.3141/1719-33.

Thorsnæs, G. *et al.* (2024) *Tønsberg*. Available at: https://snl.no/Tønsberg (Accessed: 30 April 2024).

UN council of Experts in Big Data (2023) *Use of Mobile Phone Data in Transportation*. Available at: https://unece.org/sites/default/files/2023-05/ECE-TRANS-WP6-2023-Inf-1 %28MPD Handbook%29.pdf (Accessed: 10 April 2024).

Utne, R. *et al.* (2022) *Samfunnssikkerhet*. Available at: https://www.vegvesen.no/globalassets/fag/fokusomrader/nasjonal-transportplan-ntp/2025-2036/samfunnssikkerhet.pdf (Accessed: 22 April 2024).

Valk, T., Driel, J.H. and Vos, W. (2007) 'Common characteristics of models in present-day scientific practice', *Research in Science Education*, 37(4), pp. 469–488. Available at: https://doi.org/10.1007/s11165-006-9036-3.

VanderPlas, J. (2016) *Python Data Scinece Handbook - Essential tools for working with data*. First edit. Sebastopol: O'Reilly Media Inc.

Waskom, M. *et al.* (2022) ‘mwaskom/seaborn: v0.12.2 (December 2022)’. Zenodo. Available at: https://doi.org/10.5281/zenodo.7495530.

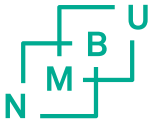
Wirth, R. and Hipp, J. (2000) ‘CRISP-DM: towards a standard process model for data mining’, *Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), pp. 29–39. Available at: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining (Accessed: 10 April 2024).

Wismans, L.J.J. *et al.* (2018) ‘Improving A Priori Demand Estimates Transport Models using Mobile Phone Data: A Rotterdam-Region Case’, *Journal of Urban Technology*, 25(2), pp. 63–83. Available at: https://doi.org/10.1080/10630732.2018.1442075.

Zhao, D. *et al.* (2010) ‘Travel mode choice modeling based on improved probabilistic neural network’, *Proceedings of the Conference on Traffic and Transportation Studies, ICTTS*, 383(April), pp. 685–695. Available at: https://doi.org/10.1061/41123(383)65.

Zhao, X. *et al.* (2020) ‘Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models’, *Travel Behaviour and Society*, 20(February), pp. 22–35. Available at: https://doi.org/10.1016/j.tbs.2020.02.003.

Zheng, Y. *et al.* (2008) ‘Understanding mobility based on GPS data’, *UbiComp 2008 - Proceedings of the 10th International Conference on Ubiquitous Computing*, (February 2014), pp. 312–321. Available at: https://doi.org/10.1145/1409635.1409677.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway