University of Tartu
Faculty of Science and Technology
Institute of Ecology and Earth Sciences
Department of Geography

Master Thesis in Geoinformatics for Urbanized Society (30 ECTS)

Detecting Public Transport Mode in The City of Tartu Using Smartphone-Based GPS Data and Machine Learning Methods

Abdelrahman Galal Elnahas

Supervisor: Dr. Siiri Silm
Co-Supervisor: Dr.Amnir Hadachi

Tartu 2021

Detecting Public Transport Mode in The City of Tartu Using Smartphone-Based GPS Data and Machine Learning Methods

Abstract

Nowadays, cities are competing between them to be more green, sustainable and intelligent, which can be summarized as being a smart city. The ease of mobility around these cities and the available transportation systems are one of the major factors that determine the urban health of these metropolis. Public transport has acquired higher interest in recent years as one of the most sustainable and environment-friendly transport modes that able to stratify the increasing demand of mobility in large cities. The ability of accurately detecting and distinguishing between transportation modes is the first required step in order to carry out any subsequent analysis for the efficiency of the existing transportation system. The aim of the study is to detect public transport use in the city of Tartu from GPS data collected using smartphone application MobilityLog. Raw GPS training data along with supervised machine learning classifiers such as k-nearest neighbours (KNN), Decision Tree (DT) and Random Forest (RF) have been utilized in order to detect the different transportation modes. The results show that Random Forest model has achieved the highest prediction score of 87.443%. After prediction phase, a downstream spatial process has been used in order to filter out wrongly predicted public transport instances using model's precision and information about public transport routes location.

Keywords: Tartu, GPS, machine learning, Random Forest, smart cities, public transport, transportation modes

CERCS code: S230 –Social geography

Ühistranspordi kasutamise tuvastamine Tartu linnas nutitelefonipõhiste GPS-andmete ja masinõppe meetoditel

Lühikokkuvõte

Tänapäeval konkureerivad linnad omavahel, et olla rohelisemad, jätkusuutlikumad ja intelligentsemad, mida võib kokkuvõtlikult nimetada targaks linnaks. Liikumise lihtsus ja olemasolevad transpordisüsteemid on ühed peamised tegurid, mis määravad linna heaolu. Viimastel aastatel on ühistransport muutunud üha olulisemaks kui üks säästlikumaid ja keskkonnasõbralikumaid transpordiliike, mis võimaldab eristada üha suurenevat liikuvuse nõudlust suurlinnades. Võimekus täpselt tuvastada ja eristada erinevaid transpordiliike on esimene vajalik samm olemasoleva transpordisüsteemi tõhususe parandamiseks. Käesoleva magistritöö eesmärk on määrata ühistranspordi kasutamine Tartus GPS andmete põhjal, mis on kogutud rakendusega MobilityLog. Töös on kasutatud GPS-toorandmeid ja masinõppe meetodeid k-lähimad naabrid (*K-nearest neighbours* – KNN), otsustuspuu (*Decision Tree* – DT) ja otsustusmets (*Random Forest* – RF), et tuvastada erinevaid transpordiliike. Tulemused näitavad, et RF mudel saavutas kõrgeima prognoosiskoori 87.443 protsenti. Ennustusfaasile järgnevalt on filtreeritud valesti ennustatud ühistranspordi juhtumid, kasutades selleks mudeli täpsust ja teavet ühistranspordi marsruutide asukoha kohta.

Märksõnad: Tartu linn, GPS, masinõpe, otsustusmets, ühistransport, targad linnad transpordiliigid

CERCS kood: S230 –Sotsiaalgeograafia

Table of contents

1 Introduction

Throughout history, advances in transportation methods have been associated with the steps taken towards human civilization. The need to build large cities and connect between them in order to exchange goods and people, in addition to military considerations at war times, have been the most important motivations that contributed to the transportation revolution (Nolan, 2009). Nowadays, half of the world's population lives in large cities (Knupfer, Pokotilo, & Woetzel, 2018).In order to improve the living conditions for their citizens and their economic development, large cities compete between them to be more green, sustainable and intelligent which can be summarized as being a smart city (Bamwesigye & Hlavackova, 2019). According to the International Telecommunication Union (ITU) "*A smart sustainable city (SSC) is an innovative city that uses information and communication technologies (ICTs) and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social and environmental aspects*" ((ITU-T), 2014). The ease of mobility around these cities and the available transportation systems are among the major factors that determine the urban health of these metropolises (Knupfer, Pokotilo, & Woetzel, 2018).

Nowadays, we live in a big data age, where vast amounts of data are generated every second. Statistics show that in 2010, there were five billion mobile phones in use (Manyika, et al., 2011). Most people are using their phones to tweet and post social media updates. Many sensors and cameras are installed in streets and squares capturing traffic information and people's movements. Governmental organizations utilize computer systems to store citizens' data and enable them to finish the paperwork in a short time. Companies and organizations are storing tremendous amounts of data about their customers and suppliers in addition to financial and human resource information about their employees. There is a potential to use big data to improve many aspects in different fields of life. One of these fields is to use big data in the urban development context to help make cities smarter and more organized. As per the previous definition, smart cities use data collected to be more green, clean and efficient. Big data also helps decision-makers and city planners make the right decisions and actions (Steenbruggen, Tranos, & Nijkamp, 2014), which will finally be reflected on citizens' happiness and how they enjoy their cities. Smart transportation is one of the main keys to building smarter cities, and leveraging big data capabilities can provide better insights and a deep understanding of people's travel patterns and modes (Wang, He, & Leung, 2017).

However, collecting big data is not enough as data needs to be analyzed using tools and algorithms in order to extract meaningful information and insights (Sivarajah, Kamal, Irani, & Weerakkody, 2017).Machine learning (ML) methods are at the core of transforming big data into useful information that can help decision-makers and smart cities planners. This is due to its ability to learn from large magnitudes of data and detect the patterns among them (L'Heureux, Grolinger, El Yamany, & Capretz, 2017).

Public transport has acquired a higher interest in recent years (Saif, Maghrour Zefreh, & Torok, 2018) as one of the most sustainable and environment-friendly transport modes that able to stratify the increasing demand for mobility in large cities (Elias & Shiftan, 2012). The city of Tartu has a modern public transport system that is formed only of buses. The system has 15 bus routes and is served by 64 buses (Inner City Bus Transportation, 2021), however, the larger share of Tartu population has only medium access to the public transport system (Dijkstra & Poelman, 2015). Because of the need to design a reliable public transport system in order to plan a smart, sustainable city and improve the quality of life for city citizens, and in the light of the need of improvement of population accessibility to the current bus systems in the city of Tartu, **the main focus of this master thesis is to use data collected using MobilityLog smartphone application and machine learning (ML) algorithms along with the spatial data of the public transport system like the locations of bus routes in order to build a model that able to detect public transport use in the city of Tartu**. The ability to detect and distinguish between transportation modes is the first step to carry on any subsequent analysis of the efficiency of the existing transportation system, which leads to robust enhancement plans. In order to achieve the thesis goal, the following research questions were framed:

1. What is/are the most important mobility feature(s) that can be used to detect public transport mobility mode?

2. What is the machine learning algorithm that yields the highest accuracy in detecting public transport mobility mode?

3. How the mix between machine learning algorithms and the classical geographical analysis can improve the accuracy of the models used to detect public transport mobility mode?

This master's thesis is structured in four main divisions. The first part is a theoretical overview in chapter two, followed by the data and methods used to conduct the analysis in chapter three. Chapter four includes the results, while the chapter five is the discussion and conclusion.

2 Theoretical overview

2.1 Public transport systems

A sustainable transport system aims to provide an efficient and time-saving way to move goods and citizens around smart cities while keeping pollution and congestion levels low. It also provides a reasonably free and green area for pedestrians and cyclists to enjoy movement away from cars (Marek, Daria, & Anna, 2020). Some researches show that one of the most essential characteristics of a sustainable transport system is to follow the "*sustainable transport pyramid*" shown in figure 1, where the most commutes are done using walking followed by cycles. Private cars come at the bottom of the most used transportation modes (Scotland, 2020). This order is based on the fact that walking and cycling are the healthiest and most environment-friendly modes of transportation (Behrendt, 2016).

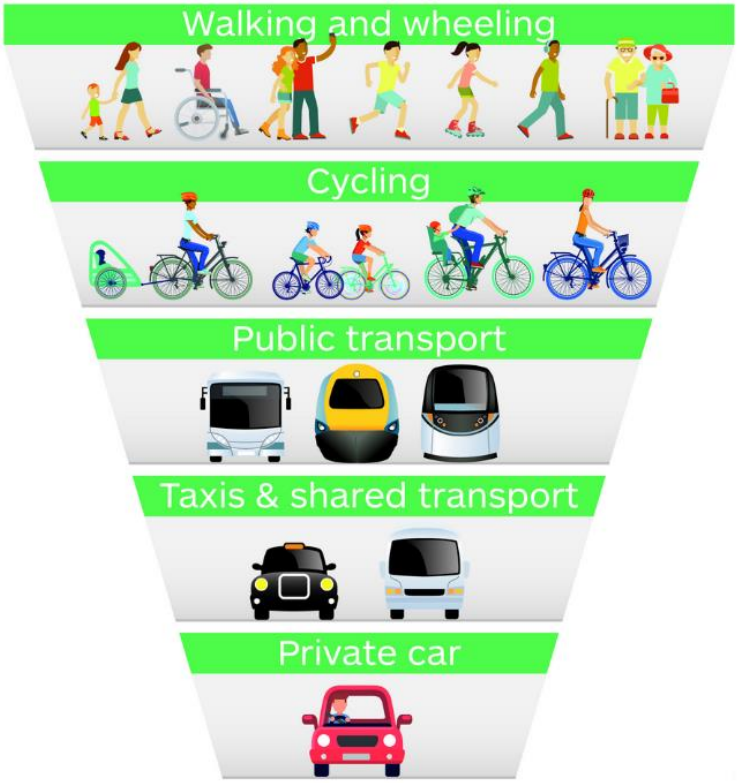

Figure 1. *Sustainable transport pyramid*
Source: (Scotland, 2020)

As per the United Nations (UN), it is expected that 68% of the world population will live in large cities by 2050, and by 2030, the world may have 43 megacities with a population of over 10 million for each (Nations, 2018). In order to satisfy the need to commute in these large cities with that huge population, a high-quality public transport system is needed. Public transport is preferred over private methods like cars for many reasons such as reducing air pollution, noise, traffic congestion, and safety (Schmöcker,

Michael, & Lam, 2003). Among all the available public transportation modes available worldwide, the statistics show that buses are the most popular used mode with 59.4% of the total usage even in cities with large rail networks like London, Sao Paulo, and Mexico City (Vakula & Raviteja, 2017). The below map in figure 2 made by the European Commission shows the access to public transport in urban centers in Denmark, Sweden, Finland, and Estonia, where the size of the cycle represents the number of populations. The map illustrates that the larger portion of city of Tartu population has only medium access to the public transportation system (Dijkstra & Poelman, 2015) , that is why this master thesis is focusing on the analysis of public bus transportation system in the city of Tartu.

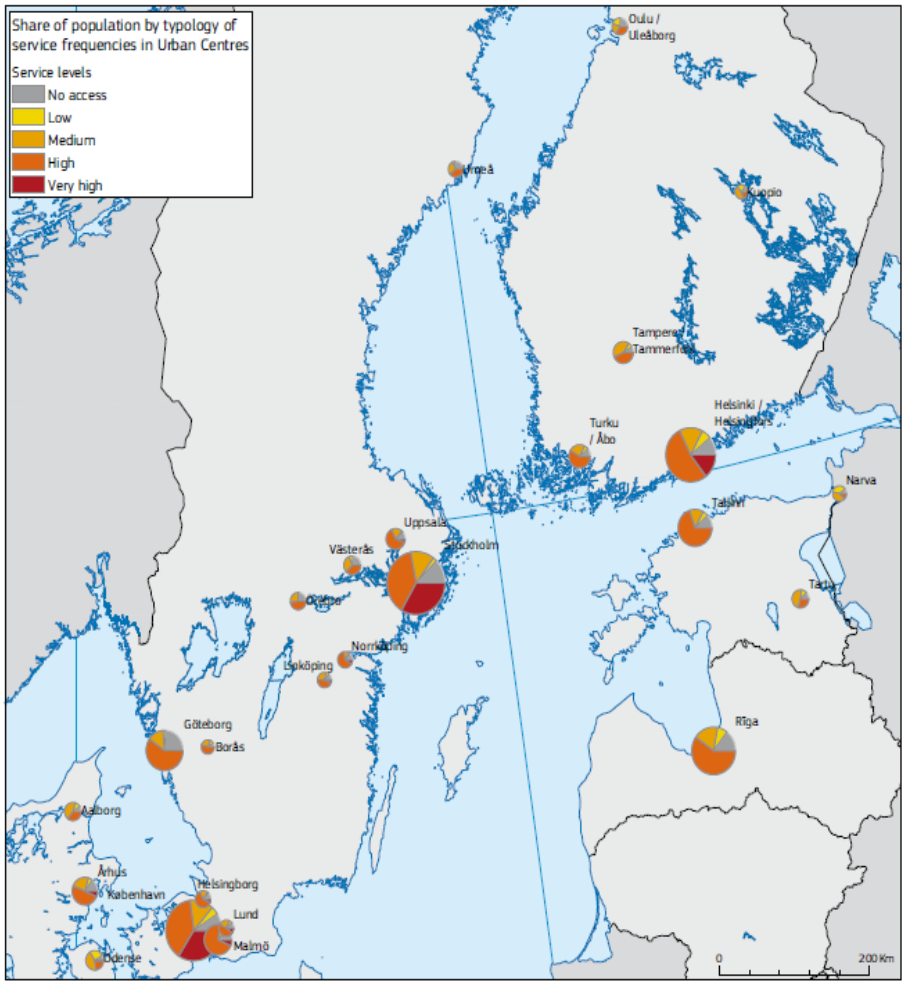

Figure 2. *Access to public transport in urban centers in Denmark, Sweden, Finland and Estonia. Source:* (Dijkstra & Poelman, 2015)

2.2 History and methods of travel data collection

The mechanism of data collection that is used to understand mobility patterns and to help network planners in order to plan cities is categorized into two main categories. The first category is the conventional data collection technique, while the second

category is the new technology collection techniques. Traditional data collection includes different kinds of road sensors, traffic cameras, and household travel surveys (HTS). Travel survey data is usually used to conduct research related to predicting human travel behaviors (Mitchell, 2014). The idea behind traditional travel surveys is that travelers document their travel history manually. Traditional travel survey data has many disadvantages like that they are hard to collect, manually generated, not structured to a standard format, and not real-time. In addition to that, data collected from traditional travel surveys have a high inaccuracy rate which makes it unreliable to determine travel start and end time in addition to travel start and end locations. Another major problem with traditional travel survey data is under-reporting trips which happen when the travelers usually consider the trip unimportant or forget to document it (Nguyen, 2015). These disadvantages and more have formulated a limitation for the research to grow further. Given the previously mentioned limitations of traditional travel surveys and the current revolution in big data methods and techniques, modern data sources can aid researchers concerned about studying travel behaviors (Wang, He, & Leung, 2017). Social media data and call detail records (CDR) are examples of using data to study user movements and locations. Global Positioning System (GPS) data is one of the widely used sources to collect information about travel behaviors (Nguyen, 2015) which is why it will be the main source of data used to conduct this analysis.

Travel surveys have been conducted in many ways over history; in the beginning, traditional travel surveys took place in the U.S.A where some interviewers visit the participant's home to have face-to-face interviews with him and manually document the response of the participant about his trip. The drawback of that method is obvious since it is costly, time-consuming, manual, and depends on the participant's memory at the time of the visit. This way has developed to a mail-out/mail-back survey method where participants receive a document by mail, fill the necessary travel information and send it back by mail. Since the late 1990s, when GPS technology became available for public use, GPS-based surveys started where participants used dedicated GPS devices to record their movement. Eventually, after the widespread of smartphones and their applications, the smartphone-based GPS survey is the most common method to collect GPS data nowadays (Shen & Stopher, 2014). Due to the importance of GPS as a source of data for travel surveys, the following section has been devoted to illustrating the GPS history and main strength points as a robust and reliable way to study and analyze travel patterns.

2.3 GPS history and advantages

The history of the Global Positioning System (GPS) backed to the 1970s when it was initially designed by the U.S. Department of Defense to be used for navigation purposes for the U.S.A military. Then it became fully operational in 1995. After GPS technology became available for non-military use and in 1996, the first trial had taken place to collect GPS data for transportation planning uses sponsored by the Federal Highway Administration (FHWA) in Lexington, Kentucky (Wolf, et al., 2014). Due to the

increasing advances in network connections, hardware capabilities, and storage capacities of the devices in addition to long battery life time, the use of GPS technology has become increasingly popular within the transportation research field. In addition to that, GPS data showed higher accuracy than traditional travel surveys with respect to the low error rates and highly precise determination of location, speed, and time of various trips (Nguyen, 2015).

According to Statista (O'Dea, 2019), 80% of Estonia's population has a smartphone that enables the users to get real-time location-based information. Smartphones usually contain embedded GPS and accelerometers that can generate data at a frequency of one point per second. In addition to that, smartphones enable users to download and install custom applications from application stores. These applications and with the support of the integrated voice recorder, keyboard and camera, give the phone owner the ability to record and document his movement and generate active travel surveys which include location, time, and even pictures of the trip (Wolf, et al., 2014). Modern programming languages and mobile operating systems like Android can also interact with embedded GPS devices, which allows the software developers to develop various smart applications that can record their movements and leverage embedded GPS devices (Komal, Khivsara, & Bramhecha, 2017). The idea of using devices owned by participants like smartphones during conducting travel surveys can facilitate and solve many problems when implementing travel surveys. These problems are the need to purchase new GPS devices for each participant and the cost of losing the device itself. In addition to that, once smartphones are connected to the Internet, it is faster to transfer data to the end databases (Bricka & Murakami, 2012). Another critical point is that travelers usually use their phones during trips, and they are not required to do any extra steps to record their movements when they are using cell phones.

Call Detail Records (CDR) data is sometimes used instead of GPS data to conduct travel pattern analysis and transportation mode detection studies as they share the same advantages with GPS data in addition to the fact that they could cover larger sample sizes. However, this advantage over GPS comes with the cost of the inaccuracy of spatial precision as CDR data contains the location of the tower that handles the call and provides the service, not the cell phone location itself (Oliver, Rein, Erki, & Robert, 2017). This disadvantage makes it impossible to use spatial data representing the city public transportation system to verify and compare the results with the machine learning model that this analysis is based on. Due to this limitation and the fact that GPS data is available for Tartu city, the decision was made to carry on the analysis using GPS data.

Since GPS devices can generate data with a frequency of one point per second, this can lead to a massive amount of data containing millions of records in a short time. Data with such high volumes requires a huge storage space, and they are hard to process; that is why the data generation frequency can be reduced on the GPS devices to be a point

every three seconds or a point every five seconds. The point generation frequency reduction can lower the amount of data significantly and save the GPS device's battery from draining after a short time of usage (Shen & Stopher, 2014).

GPS data processing usually starts with a process to divide the GPS points stream into sub-trips called segments where each segment has a different travel mode. Trip segments can be calculated using basic trip details and travel mode transitions. Basic trip details are the start point, the endpoint, and trip speed, while travel mode transition is the point where the travel mode changes from motorized to non-motorized or vice versa. Once trip segments are identified, they can be classified into different travel modes based on speed and the acceleration of the movement during the trip (Wolf, et al., 2014).

2.4 Machine learning

Machine learning (ML) can have multiple definitions, one of these definitions is "*Machine Learning is the science (and art) of programming computers so they can learn from data*" (Géron, 2019). Another definition of machine learning is" *Machine Learning is a field of computer science that evolved from studying pattern recognition and computational learning theory in artificial intelligence. It is the learning and building of algorithms that can learn from and make predictions on data sets. These procedures operate by construction of a model from example inputs in order to make driven predictions or choices rather than following firm static program instructions*" (Simon, Deo, Venkatesan, & Babu, 2015). Machine learning algorithms normally use large amount of training data as their input for learning and use what they have learnt in a form of a mathematical function to classify or predict unlabled data (Simeone, 2018).

Machine learning algorithms can be categorized into five main categories, which are: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and transfer learning (Chollet, 2017). The focus will be on the first category as it is much related to the analysis.

Supervised machine learning is a kind of algorithms that use labeled data to learn in order to be able to predict classes for unlabeled data. In other words, the algorithm is fed with an input training dataset where the output is known during the learning phase, and then the model is used to predict the output for input instances with unknown output labels (Kotsiantis, 2007). One classic example of supervised machine learning algorithms is the "cat-dog" image detection experiment, where the algorithm takes an input training set as many pictures of cats and dogs, each image is labeled correctly as cat or dog. After the learning phase, the algorithm takes an input image that is not labeled. A well-trained algorithm will be able to detect the label/class of the image correctly and determine whatever the image describes a cat or a dog.

Supervised machine learning techniques are divided into two main categories, which are classification and regression. Classification is when the output that the machine learning is trying to predict is of categorical or discrete classes. On the other hand, regression machine learning models are designed to predict numerical or continues outputs (Garbade, 2018). One case example showing the difference between classification and regression problems is the weather forecast case (Figure 3). A model that can predict as an output whatever the day is hot or cold is a classification model, while the model that has it is output as the predicted numerical temperature of the day is considered as a regression model (Vieira & Paixao, 2018).

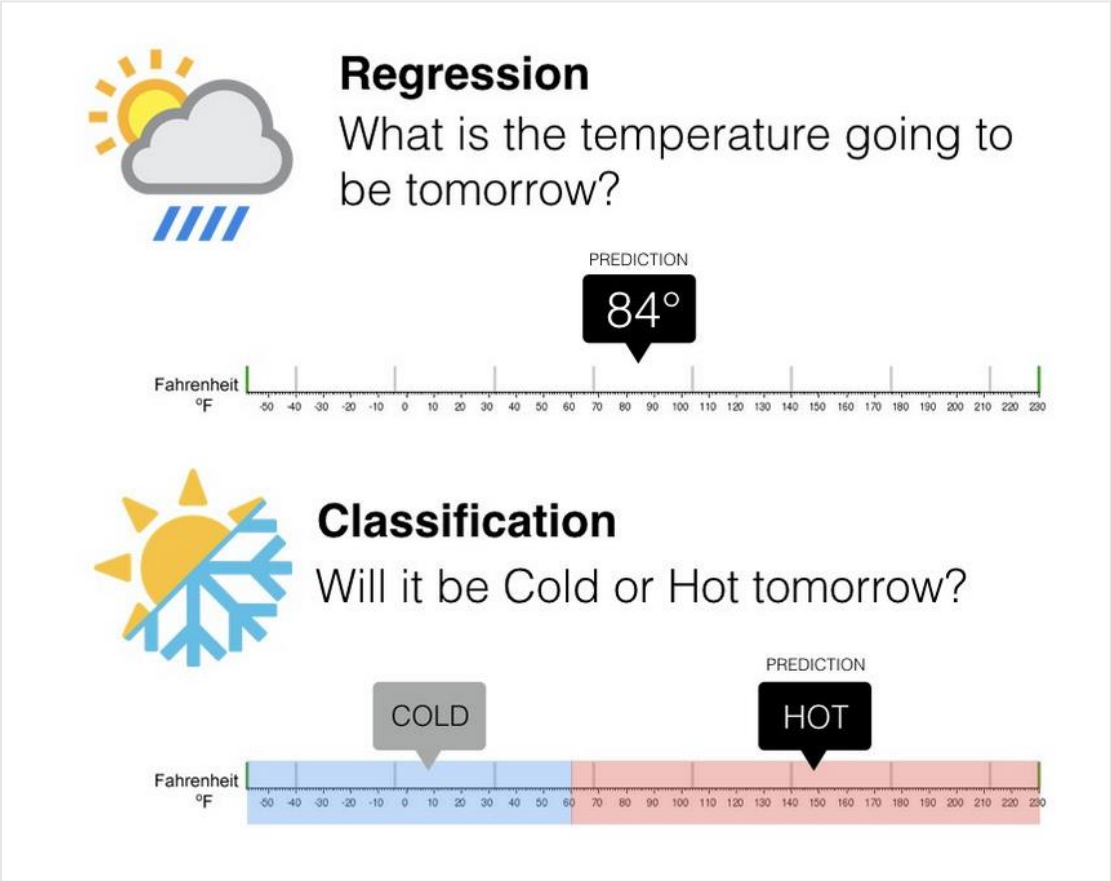

Figure 3. *Comparison between regression and classification*
Source: (Kohani, 2017)

3 Data and methods

The complete cycle to reproduce the analysis contains the following steps:

1. Data collection
2. Data cleaning and wrangling
3. Data segmentation
4. Applying machine learning algorithms
5. Measuring the accuracy of machine learning model
6. Selection of the best model
7. Applying the best model on MobilityLog unlabelled data
8. Spatial filtering of the segments wrongly predicted by the model as {*Bus*} using spatial data of the public bus system of the city of Tartu.

Figure 4 shows a summary of the workflow used in the study.

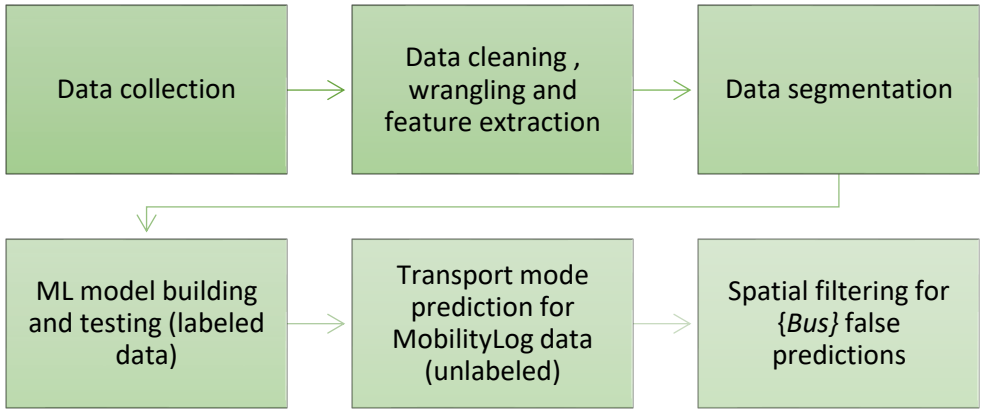

Figure 4. *Summary of the workflow used in the study*

3.1 Data sources

The main type of data used to conduct the analysis is raw GPS data. These GPS raw data are divided into two main categorize. The first category is called **labeled** data, where the transportation mode of each GPS point is known. Three labeled datasets have been used. Each separate dataset represents one of the three motorized/non-walking transportation modes that the analysis aims to predict: {*Bike*, *Bus*, *Car*}. The other category of raw GPS data is called **unlabeled** data, where the transportation mode of each GPS point is unknown. Only one unlabeled dataset has been used during this analysis. The study aims to use the three labeled datasets to train a machine learning model so that the model can predict the transportation mode of GPS points in the unlabeled dataset. In addition to raw GPS data, one spatial dataset representing Tartu city public transport routes is used. This dataset is used to verify and filter machine learning predictions. The details and description of all the data used to conduct the analysis are described below.

1. MobilityLog mobile application (unlabeled)
   This dataset includes a stream of raw GPS points with unknown labels (transportation modes). The data is stored in a computer system managed by the University of Tartu Mobility Lab and subject to continuous monitoring. The data consists of MobilityLog based location data (GPS) on the period 1.09.2019 to 30.09.2019 for 32 people containing students and staff members of the University of Tartu participating in the activity space and mobility survey of the University of Tartu Mobility Lab. I have signed an agreement to use these data for research purposes only during the period of working on the thesis. This dataset is user-based which means that it records the events/locations of MobilityLog mobile application users. Number of rows in the dataset is 1,039,908
   - Important attributes are described in table 1 and a sample of the data itself is shown in table 2.

Table 1. *Data description of MobilityLog dataset*

Attributes	Description
id	ID of the user
time_gps	Time of the event in UNIX format
time_gps_ts	Timestamp of the event in yyyy-mm-dd HH:MM:SS format
point	Coordinates of location of event in geom format
speed	Speed of the user at each point in meters/second

Table 2. *Data samples of MobilityLog dataset*

id	time_gps	time_gps_ts	point	speed
11	1.56837E+12	2019-09-13 13:56:28+03	C1E739400000A49630594D4	25.08
11	1.56784E+12	2019-09-07 11:23:46+03	6A09E38400000A6C933A84D	29.62
11	1.56768E+12	2019-09-05 13:05:18+03	429FB83A4000005E0CA2304	1.06
11	1.56837E+12	2019-09-13 13:56:15+03	DE839400000782CD6584D40	24.86
11	1.56925E+12	2019-09-23 17:48:30+03	20E61000000000C8A11BBA3	2.71
11	1.56768E+12	2019-09-05 13:05:12+03	483E97B83A4000009ECBA23	1.22
11	1.56925E+12	2019-09-23 17:48:38+03	18B224BA3A40000060AB2D3	1.4

2. City of Tartu public transport data (labeled)
   This dataset includes raw GPS points of bus trips of the city of Tartu public transport system on the period of 01.10.2020 to 31.10.2020. This data is used to represent {*Bus*} label/transportation mode when training the machine learning model. It is important to highlight that this dataset is trip-based, not user-based, which means that points represent bus trips, not user trips. Number of rows in dataset is 9,809,697
   - Important attributes are described in table 3 and a sample of the data itself is shown in table 4.

Table 3. *Data description of public transport dataset*

Attributes	Description
tripId	ID of the trip
vehicleId	ID of the vehicle
timestamp	Timestamp of the event in UNIX format
latitude	Latitude of vehicle position
longitude	Longitude of vehicle position
Trans_mode	Transportation mode for each point. All points are labeled as "Bus"

Table 4. *Data samples of public transport dataset*

tripId	Timestamp	latitude	longitude	vehicleId	Trans_mode
1011557	1602023391	58.379826	26.721577	15405	Bus
1011557	1602023380	58.379826	26.721577	15405	Bus
1011557	1602023370	58.379826	26.721577	15405	Bus
1011557	1602023361	58.379826	26.721577	15405	Bus
1011557	1602023351	58.379826	26.721577	15405	Bus
1011557	1602023339	58.379826	26.721577	15405	Bus
1011557	1602023330	58.379826	26.721577	15405	Bus

3. Bike share system data (labeled)
   This dataset includes raw GPS points of users' bike trips using the city of Tartu bike share system. Tartu Smart Bike Share is a smart bike-sharing system that serves the city of Tartu and consists of approximately 750 bikes which 250 of them are regular, and the remaining are electric bikes (Share, 2021). This dataset covers the trips in the period of 02.06.2019 to 14.06.2019 and for 3506 different users. This data is used to represent {*Bike*} label/transportation mode when training the machine learning model. It is important to highlight that this dataset is user-based. Number of rows in the dataset is 5,5881,141
   - Important attributes are described in table 5 and a sample of the data itself is shown in table 6.

Table 5. *Data description of bike share dataset*

Attributes	Description
userID	ID of the user
cyclenumber	ID of the bike
coord_time	Timestamp of the event in HH:MM:SS format
coord_date	Date of the event in yyyy/mm/dd format
latitude	Latitude of the user position
longitude	Longitude of user position
speed	Speed of the user at each point in meters/second
Trans_mode	Transportation mode for each point. All points are labeled as "Bike"

Table 6. *Data samples of bike share dataset*

cyclenumber	latitude	longitude	coord_date	coord_time	userID	Trans_mode
90417	58.38993167	26.68026833	02-06-19	11:55:57+00	28232	Bike
90417	58.38993167	26.68026833	02-06-19	11:56:02+00	28232	Bike
90417	58.38993167	26.68026833	02-06-19	11:56:07+00	28232	Bike
90417	58.38993167	26.68026833	02-06-19	11:56:12+00	28232	Bike
90417	58.38993167	26.68026833	02-06-19	11:56:17+00	28232	Bike
90417	58.38993167	26.68026833	02-06-19	11:56:22+00	28232	Bike
90417	58.38998	26.68013333	02-06-19	11:56:27+00	28232	Bike

4. Car dataset(labeled)
   This dataset includes raw GPS points of users' car trips using their own private cars. This data is obtained from two users and has been collected specifically for this study. Car usage data from anonymous users could not be obtained because they are not available/collected on the same way like bike share system data. The users have recorded their car trips using software called Komoot (Komoot, 2021). This dataset covers the trips in the period of 08.03.2021 to 14.03.2021. This data is used to represent {*Car*} label/transportation mode when training the machine learning model. It is important to highlight that this dataset is user-based. Number of rows in the dataset is 12,561
   - Important attributes are described in table 7 and a sample of the data itself is shown in table 8.

Table 7. *Data description of Car dataset*

Attributes	**Description**
userID	ID of the user
timestamp	Timestamp of the event in yyyy-mm-dd HH:MM:SS format
latitude	Latitude of the user position
longitude	Longitude of user position
Trans_mode	Transportation mode for each point. All points are labeled as "Car"

Table 8. *Data samples of Car dataset*

userID	timestamp	Latitude	longitude	Trans_mode
2	2021-03-08T15:12:58.880Z	58.368755	26.722853	Car
2	2021-03-08T15:13:54.177Z	58.368576	26.722887	Car
2	2021-03-08T15:13:57.197Z	58.368458	26.722882	Car
2	2021-03-08T15:14:10.180Z	58.368374	26.722977	Car
2	2021-03-08T15:14:12.171Z	58.368411	26.723159	Car
2	2021-03-08T15:14:14.175Z	58.368472	26.723402	Car
2	2021-03-08T15:14:16.179Z	58.368532	26.723631	Car

5. Public transportation routes
   In addition to the previous datasets, the last dataset contains the spatial aspects and attributes of the public bus transportation system of Tartu city. The data set includes coordinates of bus routes, bus stops locations, and bus route names. This dataset is used as a second layer of verification and filtering after the machine learning model detects the transportation modes. The data consist of a sequence of points through which the vehicle passes in order, so each route is represented by a sequence of points, not a single line. Information and details about this dataset can be found in (Google, 2021).Number of rows in the dataset is 1048,576
   - Important attributes are described in table 9 and a sample of the data itself is shown in table 10.

Table 9. *Data description of public transportation routes dataset*

Attributes	Description
shape_id	ID of the user
shape_pt_lat	Latitude of the user position
shape_pt_lon	Longitude of user position
shape_pt_sequence	Speed of the user at each point in meters/second

Table 10. *Data description of public transportation routes dataset*

shape_id	shape_pt_lat	shape_pt_lon	shape_pt_sequence
132	58.753472	24.9425	1
132	58.753545	24.942327	2
132	58.753743	24.942277	3
132	58.75423	24.942299	4
132	58.754435	24.942356	5
132	58.754942	24.942692	6
132	58.755266	24.942844	7

3.2 Data cleaning and wrangling

3.2.1 Data cleaning

Most datasets come dirty with many problems like missing values, duplicate values, and outliers. These problems can affect the accuracy of any analysis in a significant way (Jesmeen, et al., 2018). For this reason, many data cleansings steps have been followed to make the data cleaner before using it in the analysis. These steps are described below.

1- Each dataset has been filtered in order to exclude any GPS points that exist outside the boundaries of the city of Tartu.

2- Since datasets are coming from different sources, timestamp has been unified across all datasets to be on Estonian Time Zone.
3- Data records with missing values have been removed for all datasets.
4- Duplicated records have been removed for all datasets.
5- Data has been grouped by a user then ordered by date and timestamp to reflect each user's real movement sequence. A slightly different approach has been applied to the bus dataset since it is trip-based, not user-based. For the bus dataset, data has been grouped by trip then ordered by date and timestamp to reflect each bus's real sequence of movements.
6- All irrelevant columns have been removed from all datasets.

3.2.2 Data wrangling

Data wrangling can be defined as the process of preparing and formatting the data (Patil & Hiremath, 2018) in order to put them in a form that is usable for conducting the required analysis (Kandel, et al., 2011). This section describes the data wrangling processes that have been applied to used datasets.

1- Rows in all raw GPS datasets have been formatted as below. This data structure gives the ability to calculate time difference, distance, and speed between consecutive points.
   - *date_start: Date of GPS point*
   - *date_end: Date of the next GPS point*
   - *time_start: Time of GPS point*
   - *time_end: Time of the next GPS point*
   - *latitude_start: Latitude of GPS point*
   - *latitude_end: Latitude of next GPS point*
   - *longitude_start: Longitude of GPS point*
   - *longitude_end: Longitude of next GPS point*
   - *userID/tripID: User identifier in case of Bike/Car or Trip identifier in case of Bus of GPS point*
   - *userCheck/tripCheck: User identifier in case of Bike/Car or Trip identifier in case of Bus for the next GPS point*

2- Rows where *userID/tripID* and *userCheck/tripCheck* do not hold the same values have been removed. Change in the values between the two fields indicates the end of the GPS points stream belongs to the value of user or trip in *userID/tripID* and highlights that the next row is a start for another GPS points stream belonging to the user or trip of *userCheck/tripCheck.*

3- The time difference between each two consecutive GPS points has been calculated by subtracting the start timestamp from the end timestamp and converting the results to seconds. The following formula has been used.

Δt = (*time_end - time_start).to_seconds()* (1)

where:

- Δt *is the change in time.*
- *to_seconds() is a function to convert the time difference to seconds*

4- Distance in meters has been calculated between each two consecutive GPS points. There are many methods that can be used to calculate the distance between two GPS points, such as Haversine and Trapezoidal; however, Haversine formula tends to be more accurate for the GPS points obtained through mobile applications (Lindenberg, 2014) which is the case for the used raw GPS data used to through this analysis. The following Haversine formula has been used to calculate the distance between each two adjacent GPS points (Lindenberg, 2014):

$haversin\ \theta = sin2(\theta/2)$

$d = 2r\ sin{-}1(\sqrt{haversin(\phi 2 - \phi 1) + cos(\phi 1)cos(\phi 2)haversin(\lambda 2 - \lambda 1)})$ (2)

Where:

1- $\phi 1$ *latitude of the first point p1 or latitude_start*
2- $\lambda 1$ *longitude of the first point p1 or longitude_start*
3- $\phi 2$ *latitude of the first point p2 or latitude_end*
4- $\lambda 2$ *longitude of the first point p2 latitude_end*
5- *r is the radius of the Earth in meters*
6- *d is the distance in meters between p1 and p2*

5- Speed has been calculated between each two consecutive GPS points. The following formula has been used.

$speed = d/\Delta t$ (3)

Where: d is distance and Δt *is time difference and speed in meters/second*

3.3 Methods

3.3.1 Data segmentation

Zheng, Liu, Wang, & Xing (2008) have proposed a robust technique that can be used to better divide the GPS points streams into segments based on the change of the transportation mode. This framework will be adopted to perform segmentation for MobilityLog mobile application unlabeled dataset. In order to better understand the framework of segmentation, few definitions should be introduced:

1- **Track**: the track is a series of GPS points belonging to the same user.

2- **Trip**: Trip is a subset of the track. When the time difference between two consecutive GPS points in the same track exceeds a certain time threshold or

certain distance threshold, this indicates the end of a trip and the start of a new one.

3- **Segment**: Segment is a subset of a trip that satisfies the condition that all the GPS points included in the segment should all share the same transportation mode. From this definition, it is normal to have a single trip that contains multiple transportation modes like Car->Walk->Bus. This trip contains three segments and three different transportation modes. An example that illustrates the definition more is an example of a student who left his home and went to the university. During this trip, he walked to the bike station, took a bike to the nearest bus station and finally took a bus that dropped him at the university. In this case, all the GPS points between his home to the university represent the trip. However, this trip can be divided into three different segments labeled with {*Walk, Bike, Bus*} transportation modes. If the student had walked all the way from his home to the university, then the trip and segment are the same because there has been no change in transportation mode during the trip, and no further segmentation is needed.

4- Segments can be categorized into two categories. The first category is *non-motorized* segments which have "*Walk*" as a transportation mode, while the second category is *motorized* segments which can have any other type of transportation modes including {*Bike, Bus, Car*}.

- Segmentation of labeled datasets
  Each dataset of the three labeled datasets used in this analysis has its unique transportation mode. Therefore, no need to search for the change in transportation mode when applying the segmentation process. Since all the points in each single dataset share the same transportation mode, then trip and segment definitions are the same and it is only required to divide each of the user tracks into many trips. The trip division strategy is based on setting a time and a distance threshold. If the time or distance difference between any two consecutive GPS points in the same track have exceeded these thresholds, this is marked as a discontinuity in the user movements, and then the track is divided into different trips. The following algorithm has been used to segment both Tartu City smart shared bike system and car datasets.
  The value of time threshold has been set to 60 seconds, while the distance threshold has been set to a value equals to the time difference between each two consecutive GPS points multiplied by average speed for GPS points in each dataset. For bikes, the average speed is 15 meters/second, while for cars, the average speed is 40 meters/second.

for each_user in all_users:
current segment = start a new segment
for point in all_GPS_points_of_that_user:
if $\Delta t > t_{threshold}$ *or* $d > d_{threshold}$*:*
end the current segment
start a new segment
else:
append point to current segment

The segmentation process of the bus dataset is more straightforward since the data is trip-based in the first place, which means all the GPS points that belong to the same trip share the same trip identifier. The segmentation process then divides the track into many segments where all the points in the same segment have the same trip identifier and represent a single bus journey.

- Segmentation of unlabeled datasets
The segmentation process of the MobilityLog mobile application unlabeled dataset is more complex than the same process for a labeled dataset for two main reasons. The first reason is the fact that it contains points that can belong to many transportation modes, which was not the case when dealing with the other three datasets where all points in each dataset represent only one transportation mode. The other reason is the absence of those transportation modes in the dataset. The idea of the segmentation process is that the transition between two different motorized transportation modes usually contains a non-motorized segment in between where speed is zero or very low. The first step is to separate the GPS tracks of each user to motorized and non-motorized trips. Distinguishing between motorized and non-motorized transportation modes is not difficult because human walking speed is usually around 6 km/hour, while the speed of motorized vehicles is usually above 40 km/hour (Shen & Stopher, 2014). However, having multiple motorized transportation modes could be a problem as buses, cars, and bikes may have similar GPS characteristics (Stenneth, Yu, & Xu, 2011). Distinguishing between non-motorized segments will be carried on by machine learning models in later analysis stages. After applying this first technique, the result is a stream of congestive (*"motorized," "non-motorized*," *"motorized"*) streams. Some of these middle *non-motorized* segments can represent a real walking period; however, some of them could represent a gap in a motorized segment due to a stop for a traffic light or even a bus stop. In order to fix this problem, very short *non-motorized* segments between two *motorized* segments are removed, and the next and previous motorized segments are merged into a single segment if there is no time or distance gap between them.

- After GPS raw data segmentation, each segment is aggregated into a single row that contains the main important attributes like user or trip identifier, the number of points in the segment, total time, total distance, and more importantly, the *average speed* and *average acceleration* of the segment, in addition to the transportation mode associated with each segment in case of labeled data. The segmentation process is beneficial because it models the real problem where the analysis is concerned about the whole user trip rather than studying each individual GPS point. In addition to that, the segmentation process significantly reduces the number of records. This, in turn, helps to improve the speed performance of the subsequent machine learning algorithms and lower the computational resources needed to complete the analysis (Luna, Cano, & Ventura, 2016).

- After completing the segmentation process for each dataset separately, the three output datasets are merged into a single dataset containing segments. Each segment is labeled with one of these transportation modes {*Bike, Bus, Car*}.

3.3.2 Applying machine learning

The output of the segmentation stage is a single segments dataset. This dataset is used as an input for the next stage, which is the stage of building the machine learning model. Since machine learning algorithms are categorized into many categories and even each category can be divided into other subcategories, it would be useful to accurately define the computational problem that the analysis aims to solve. This definition will help to choose the appropriate methods. The definition of the problem is "*the need to build a machine learning model using a labeled dataset as an input training set in order to predict the transportation modes of another unlabeled dataset accurately. The transportation modes that the machine learning model would predict could be either {Bike, Bus, Car}".* Since the required machine learning techniques use a labeled training set during the learning phase, supervised machine learning methods are the best algorithms to use. In addition to that, the supervised machine learning models try to predict categorical values, not continuous ones, which means classification supervised machine learning methods are the most appropriate method to solve that computational problem. Since supervised machine learning classifiers depend on feature inputs to build models, then the model output and prediction accuracy will differ. The analysis will compare between the prediction accuracy of different supervised machine learning models when using only the average speed or average acceleration of each segment as an input feature and the prediction accuracy of the same models but when they are trained using both average speed and average acceleration of each segment as input features.

When supervised machine learning algorithms are used, a typical workflow is to split the training set into two parts called "learning set" and "testing set". The learning set is used as an input to machine learning to build the model, while the testing set is used to measure the accuracy of the model predictions by comparing them with the true labels in that set. The ratio between the size of the "learning set" and "testing set" can affect the model performance and accuracy. In general, there is a positive correlation between the size of the learning dataset and the accuracy of the model (Medar, Rajpurohit, & Rashmi, 2017). For this analysis, the data has been split, so the ratio of the learning dataset to the testing dataset is .7:.3.This ratio ensures that the model has enough large amount of data during both the learning and testing phases.

Another aspect to consider with the input learning dataset is the unequal distribution of labels. This happens when one label exists more frequently than other labels in a learning data. In this case, the dataset is considered to be an *imbalanced dataset* (Chawla, 2005). Feeding supervised machine learning models with an imbalanced learning dataset can cause a negative impact on the accuracy of classification (Feng, Huang, & Ren, 2018).The reason is that most supervised machine learning classifiers are designed to maximize the overall prediction accuracy. As a result, they tend to classify all the data into the majority label (Kotsiantis, Kanellopoulos, & Pintelas, 2005). In the balanced dataset, the instances labeled as {*Bike, Bus*} will be equally represented, while in the imbalanced dataset, the instances labeled as {*Bike}* will be represented more frequently those labeled as *{Bus}*.

In order to complete these analyses, three different supervised machine learning classifiers are used. The used classifiers are the *k-nearest neighbors* (KNN), *Decision tree,* and *Random forest* classifiers. These classifiers have been chosen as they are among the most used machine learning classifiers (Minastireanu & Mesnita, 2019). A brief description of each classifier is below.

- *k-nearest neighbours (KNN)*
  k-nearest neighbours supervised machine learning classifier, which is simple and accurate. The classification is based on the majority voting of labels for the k nearest instances of the target (Guo, Wang, Bell, & Bi, 2004). In other words, in order to predict a label of an instance, the algorithm selects the nearest k records from that instance and determines its label based on the major label of these selected k records. The accuracy of the prediction of the k-nearest neighbors is mainly determined by the value of k (Imandous & Bolandraftar, 2013). During this analysis, the k-nearest neighbors will be applied with different values of k=1,*k=3, k=5, k=7,* and *k=9.* Since k-nearest neighbors is a voting algorithm, it is better to choose k values as odd numbers to ensure that there will be no instance that has equal votes representing different labels (Hassanat, Abbadi, & Alhasanat , 2014).

- *Decision tree*
  Decision tree is a supervised machine learning classifier that is similar to the process that a human may follow to make a decision. It starts with a node that tests some criteria depending on the input features. The output of these nodes is a binary decision that branches to two more additional nodes. The splitting criteria continue until a prediction is made (Patel & Prajapati, 2018).

- *Random forest*
  Random forest is a supervised machine learning classifier that works in a way similar to a decision tree classifier. However, instead of making a prediction based on only one single tree, the algorithm splits learning data randomly then it constructs *n* number of decision trees. After this process, the model's prediction is made based on the majority voting of these trees (Ren, Cheng, & Han, 2017). Random forest classifier has proved to be an accurate and efficient algorithm (Ali, Khan, Ahmad, & Maqsood, 2012). It has many applications in different fields like traffic and transport planning, ecology, astronomy, and agriculture (Fawagreh, Gaber , & Elyan, 2014).

During the analysis, the three mentioned machine learning models are trained using two datasets: imbalanced and balanced. With each model and dataset pair, the speed feature of learning segments is used as an input feature in one round. In the second round acceleration feature of learning, segments are used as an input feature, and in the last round, both speed and acceleration features are used during the training process. After training all the models using the previous setup, trained models are applied to the same test dataset and prediction *accuracy* is measured as the percentage of correctly predicted instances to the total number of instances in the test dataset (Liu, Zhou, Wen, & Tang, 2014). In order to find the most fit model and examine the impact of using imbalanced data as a training dataset, the model with the highest accuracy from the two setups is selected. These two models are examined in terms of two measures called *recall* and *precision*. Recall is the ratio of true positive predictions to the total of true positive and false negative predictions for each class, while precision is the ratio of true positive predictions to the total of true positive and false positive predictions for each class (Goutte & Gaussier, 2005). Based on these criteria , the final model is selected and eventually applied to motorized MobilityLog data segments to predict their classes. The selected model indicates what is the important feature(s) to include during the learning phase, whether speed or acceleration alone or both.

3.3.3 Spatial filtering

The focus on this step is directed to MobilityLog segments that the chosen model has classified to be of class {*Bus*} as it is the transportation mode of interest. A spatial method is used as a downstream process after applying the selected machine learning algorithm on MobilityLog motorized segments to filter out {*Bus*} false positive instances (segments that the model predicts to be of class {*Bus}* while actually, they are

not). This step helps to increase the model precision for the {*Bus*} class. The idea is to calculate the percentage of GPS points that are geographically located *within* Tartu city bus routes to the total number of points in each segment and name this metric as "*within bus route ratio*" (Figure 5). The second step is to order the segments in a descending way based on their "*within bus route ratio"*. The precision of the model indicates the percentage of these segments that actually are {*Bus*} segments. A percentage of segments equals to the model precision starting from those ones with higher "*within bus route ratio"* are accepted to be of the {*Bus*} class while the others are filtered out as false positive cases.

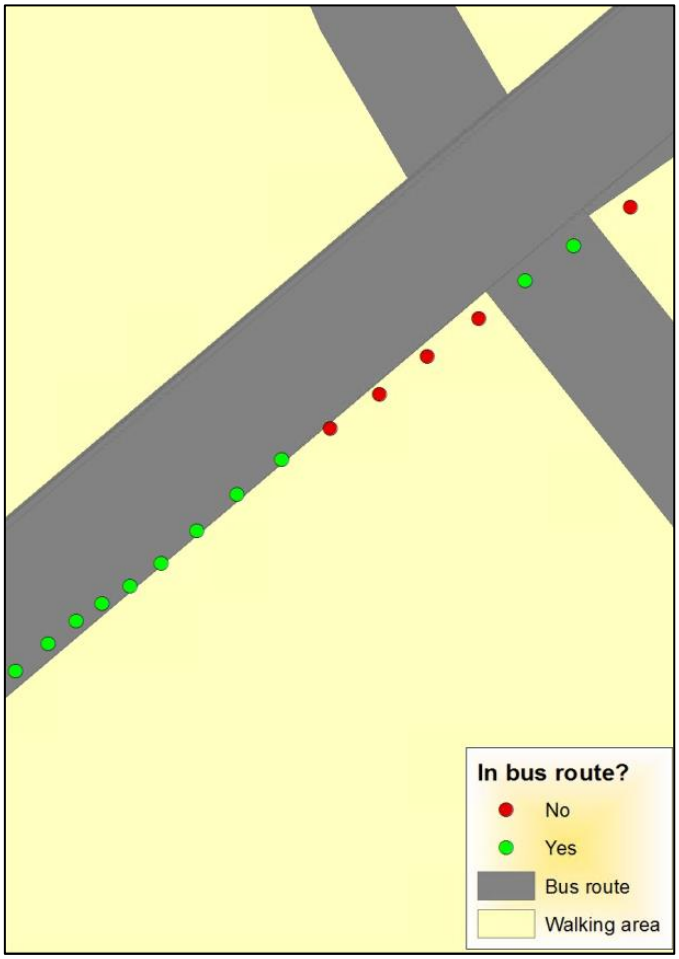

Figure 5. *within bus route ratio is the ratio between green points to the total points in the segment*

4 Results

4.1 Data segmentation

The results of the raw GPS points segmentation process are illustrated in table 11. The results show that a total of 66,048,259 GPS points have been segmented to 616,925 segments with a total reduction ratio equals 99.07%. The individual reduction ratios for each dataset have been ranged between 98.58% for the MobilityLog dataset and 99.77% for the car dataset. The segmentation process results are four datasets called MobilityLog, Bike, Bus, and Car datasets for the rest of the analysis. Each dataset contains a number of segments with two features called *Speed* and *Acceleration.* These two features represent the average speed and the average acceleration for all the points in each segment. The 4,888 MobilityLog segments are divided as 1,234 {*non-motorized/Walk*} segments and 3,654 {*motorized*} segments.

Table 11. *Results of the segmentation process*

	Raw GPS points	Segments	Reduction %
MobilityLog	344,860	4,888	98.58%
public bus system	9,809,697	33,902	99.65%
shared bike system	5,5881,141	578,106	98.97%
Car dataset	12,561	29	99.77%
Total	66,048,259	616,925	99.07%

The mean and standard deviation of *Speed* and *Acceleration* features for each segmented dataset are shown in table 12. Car segments have the highest average *Speed* with 10.35 meters/second, while Bike segments have the lowest average *Speed* with 3.17 meters/second. For the *Acceleration*, Bus segments have the highest average *Acceleration* with 0.0021 meters/second^2, while the lowest average Acceleration is represented by the Bike dataset with -0.0082 meter/second^2.

Table 12. *Mean and Standard deviation of Speed and Acceleration feature*

	Mean Speed	Std of Speed	Mean Acceleration	Std of Acceleration
Bike	3.17	1.24	-0.0082	0.023
Bus	4.75	1.11	0.0021	0.004
Car	10.35	3.4	-0.0141	0.077

Figure 6 shows that the *Speed* feature of Car segments has the highest standard deviation between all the other datasets with 3.4 meters/second. In contrast, Bus segments have the lowest *Speed* standard deviation with 1.11 meter/second.

For the standard deviation of the *Acceleration* feature, Car segments ranked first with the highest standard deviation of 0.077 meters/second^2. In contrast, Bus segments ranked last with .004 meter/second^2.Figure 7 shows the difference in *Acceleration* standard deviation among the three datasets.

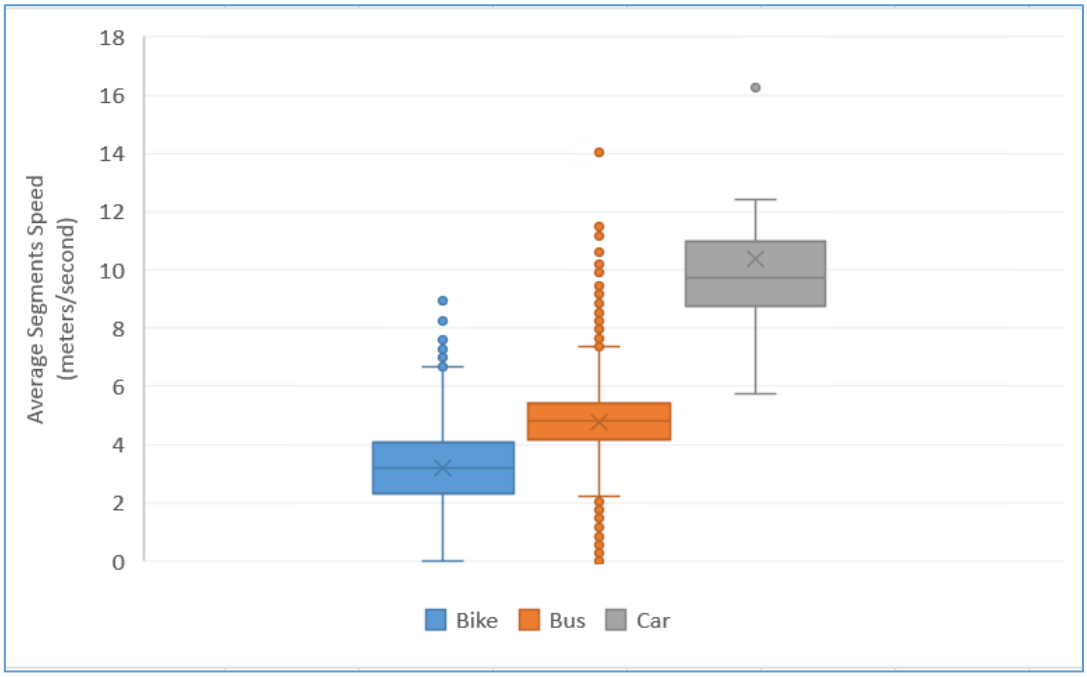

Figure 6. *Distribution of Speed feature for the different transportation mode*

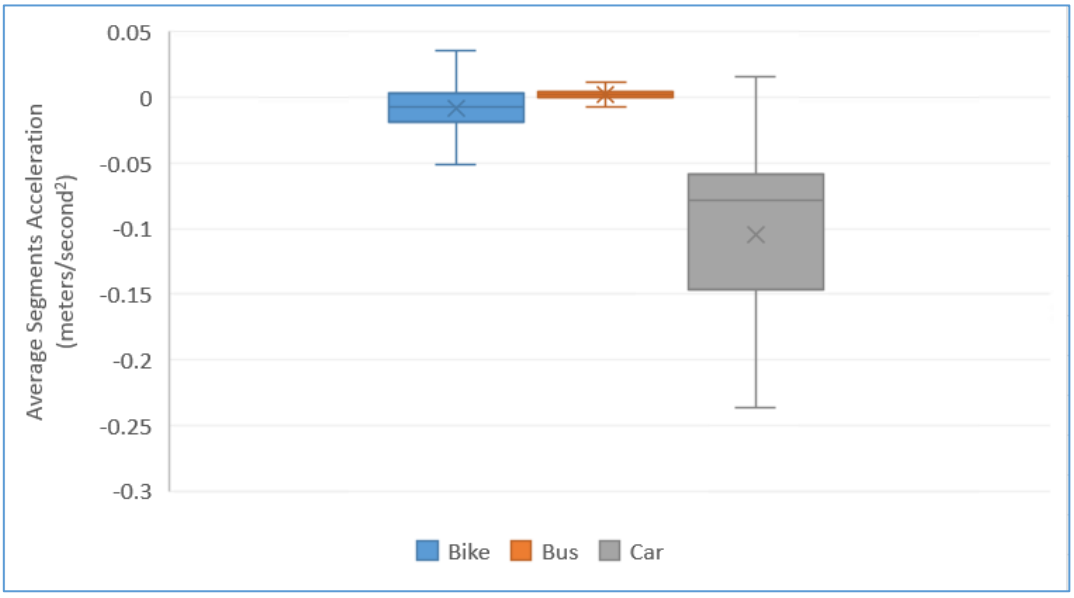

Figure 7. *Distribution of Acceleration feature for the different transportation mode*

4.2 Comparison of machine learning algorithms and mobility features

This section describes the prediction accuracy results after applying three different machine learning models (KNN, Decision tree, and Random forest) on the same test dataset. These models have been trained using two different learning sets (Imbalanced,

Balanced) using *Speed* feature in one round, *Acceleration* feature in a second round, and both *Speed* and *Acceleration* features at the last round.

4.2.1 k-nearest neighbours (KNN)

The results of k-nearest neighbours (KNN) are described in table 13. Five different values of k have been used (k=1, k=3, k=5, k=7, k=9) to train the model. The results can be summarized as below.

- The accuracy percentage of all the rounds ranged between 67.644% and 95.296%.
- In both cases of models trained using imbalanced or balanced datasets, using both *Speed* and *Acceleration* features of the segments in the training process has yielded higher accuracy than using only one feature at a time.
- For all the cases, models trained using an imbalanced dataset have resulted in higher accuracy than the corresponding models trained using a balanced dataset.
- Increasing the value of K does not increase the accuracy in all cases.
- After this round of testing, the model with the highest accuracy in the imbalanced category is KNN with K=9 and accuracy = 95.296%, while the model with the highest accuracy in the balanced category is KNN with K=5 and accuracy = 85.518%.

Table 13. *Results of k-nearest neighbours model. The yellow marker shows the highest score in the Imbalanced category, while the green marker shows the highest score in the Balanced category*

	Imbalanced			**Balanced**		
	Speed	Acceleration	Both	Speed	Acceleration	Both
K=1	91.018%	90.717%	93.8%	67.644%	74.046%	84.306%
K=3	93.483%	92.887%	94.81%	71.044%	74.576%	85.27%
K=5	94.050%	93.633%	95.1%	72.062%	74.853%	85.518%
K=7	94.243%	94.027%	95.256%	72.676%	74.538%	85.144%
K=9	94.375%	94.187%	95.296%	73.063%	74.512%	84.953%

Figure 8 shows the decision boundaries of a KNN model trained using imbalanced data, both *Speed* and *Acceleration* as features and a value of K=9. This model has a prediction accuracy of 95.296%. Different colours represent the boundaries of each transportation mode. The model places each unlabelled segment in this grid using its *Speed* and *Acceleration* features, then predicts the transportation mode corresponding to the area where the segment is located. For example, if for one segment the *Speed* feature value is 5.5 meters/second and the *Acceleration* feature value is -0.3 meters/second^2, then the segment is located in the blue area, and then the model predicts its transportation mode to be *Bike* and so on for any unlabelled segment. The model mainly classifies segments

based on their *Speed* feature with this approximate range of [0 - 7.5] meters/second to be classified as *Bikes*, [7.5 - 10.5] to be classified as *Bus* and any higher speed is classified as *Car*. *Acceleration* feature is then used to differentiate for places where an overlap exists.

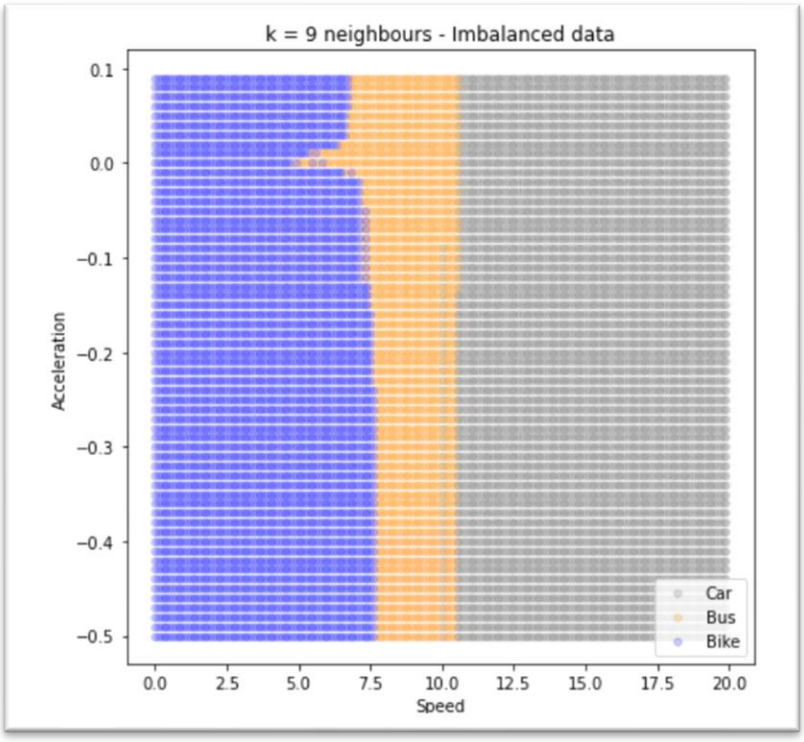

Figure 8. *KNN decision boundaries for K=9 and Speed and Acceleration as input features*

4.2.2 Decision tree (DT)

The results of the decision tree (DT) model are described in table 14. The results can be summarized below.

- The accuracy percentage of all the rounds of the decision tree model ranged between 67.633% and 93.723%.
- In both cases of models trained using imbalanced or balanced datasets, using both *Speed* and *Acceleration* features of the segments in the training process has yielded higher accuracy than using only one feature at a time.
- For all the cases, models trained using an imbalanced dataset have resulted in higher accuracy than the corresponding models trained using a balanced dataset.
- After this round of testing, the results have not been changed. The model with the highest accuracy in the imbalanced category is still KNN with K=9 and accuracy = 95.296%, while the model with the highest accuracy in the balanced category is still KNN with K=5 and accuracy = 85.518%, which

means that the accuracy of KNN model exceeds the accuracy of the decision tree model.

Table 14. *Results of the Decision Tree model, no markers as the model has not shown any improvement over the previous round*

Imbalanced			**Balanced**		
Speed	Acceleration	Both	Speed	Acceleration	Both
91.057%	92.894%	93.723%	67.663%	74.857%	84.925%

Figure 9 shows part of a decision tree model trained using imbalanced data and has a prediction accuracy of 95.2%. The decision is started from the root node, then it splits left if the test result is *True* or right if the test result is *False*. Each colour is corresponding to a certain class, the orange colour represents *Bike* transportation mode while the green colour represents *Bus* transportation mode. The *samples* value in each node represents the percentage of samples in each stage. The *values* list in each node contains three values corresponding to the probability of prediction for each class of the three classes. The node is coloured according to the class with the highest probability while the colour intensity of each node is probational for the probability value. Higher probability leads to more intense colour.

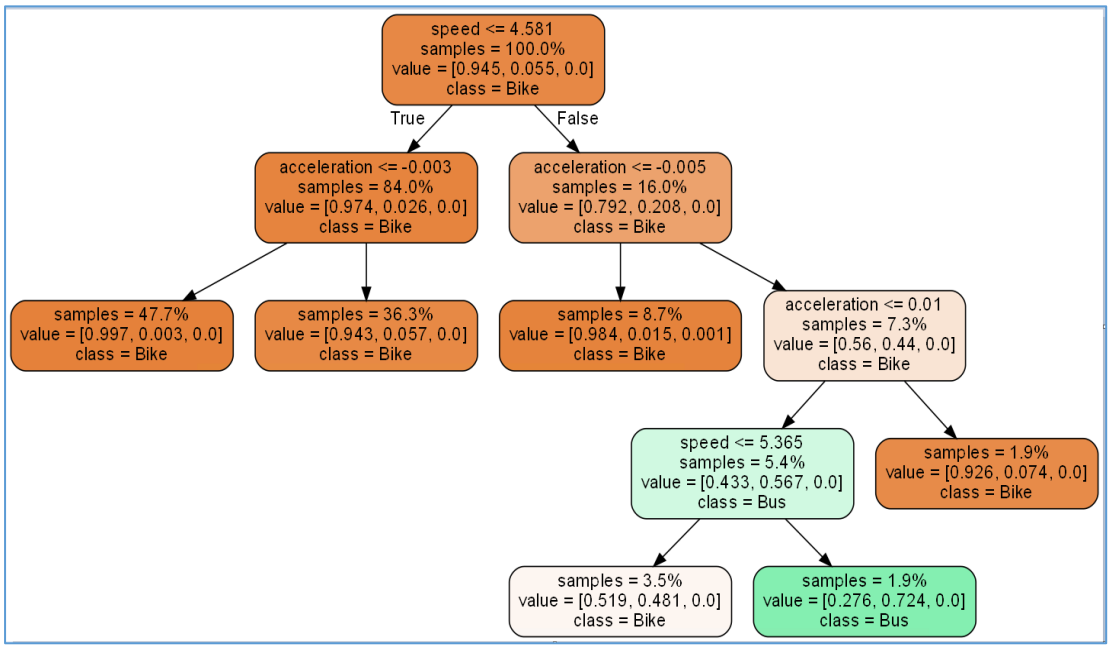

Figure 9. *Part of a Decision Tree illustration for a model trained with Speed and Acceleration as input features*

4.2.3 Random forest (RF)

The results of random forest (RF) are described in table 15. Five different values of n (number of trees) have been used (n=5, n=10, n=20, n=30, n=40) to train the model. The results can be summarized as below.

- The accuracy percentage of all the rounds ranged between 68.041% and 95.113%.
- In both cases of models trained using imbalanced or balanced datasets, using both *Speed* and *Acceleration* features of the segments in the training process has yielded higher accuracy than using only one feature at a time.
- For all the cases, models trained using an imbalanced dataset have resulted in higher accuracy than the corresponding models trained using a balanced dataset.
- Increasing the value of *n* does not increase the accuracy in all cases.
- After this round of testing, the model with the highest accuracy in the imbalanced category is KNN with K=9 and accuracy = 95.296%. In comparison, the model with the highest accuracy in the balanced category is now random forest with n=10 and accuracy = 87.443% replacing KNN with K=5 and accuracy = 85.518%.

Table 15. *Results of Random Forest model, the green marker indicates that this model has the highest score in balanced category after this round*

	Imbalanced			**Balanced**		
	Speed	Acceleration	Both	Speed	Acceleration	Both
n=5	91.47%	92.145%	94.732%	68.133%	74.119%	86.253%
n=10	91.774%	92.352%	94.999%	69.565%	74.753%	87.443%
n=20	91.431%	92.375%	95.058%	68.454%	74.269%	87.038%
n=30	91.285%	92.394%	95.102%	68.041%	74.152%	86.853%
n=40	91.186%	92.403%	95.113%	68.879%	74.031%	86.777%

4.2.4 Model choice

The result of all rounds of model training and testing is two models that scored the highest prediction accuracy in one of each category. The first model is a 95.296% accurate KNN model trained with imbalanced data using *Speed* and *Acceleration* features and has K=9, while the second model is a 87.443% accurate random forest model trained with balanced data using *Speed* and *Acceleration* features and has n=10.A confusion matrix is used in order to compare the two models and choose the best among them in terms of accuracy, recall, and precision. A confusion matrix is a table used to assess the prediction performance of a machine learning classifier by comparing the actual classes of test instances versus the model predicted classes for the same instances (Ting, 2017). The diagonal cells of the confusion matrix represent the number of instances that have been correctly predicted by the model as their actual class matches the model predicted class, while the other cells of the matrix represent the wrong prediction in each class (Mo, et al., 2020).

The confusion matrix of the first model (KNN) is illustrated in table 16. It shows that the model has a 98.4% recall for the instances with {*Bike}* as their actual label, which

means that the model is able to correctly predict 98.4% of the actual bike instances. On the other hand, the model has a low recall for the other two classes. The model is able to correctly predict only 42.1% of instances actually labeled as {*Bus*}, while 5,885 instances which actually labeled as {*Bus*} are predicted to be {*Bike*} and hence considered as false negatives. The same applies to the instance with the {*Car*} label as the model correctly predicts only 33.33% of them.

Table 16. *Confusion matrix of KNN model trained with imbalanced data*

		Predicted				
		Bike	*Bus*	*Car*	**Total**	**Recall**
Actual	*Bike*	170,689	2,745	0	**173,434**	98.4%
	Bus	5,885	4,285	2	**10,172**	42.1%
	Car	1	3	2	**6**	33.33%
	Total	**176,575**	**7,033**	**4**	**183,612**	
	Precision	96.66%	60.92%	50%		

Table 17. *Confusion matrix of random forest model trained with balanced data*

		Predicted				
		Bike	*Bus*	*Car*	**Total**	**Recall**
Actual	*Bike*	151,629	21,803	2	**173,434**	87.42%
	Bus	1,248	8,923	1	**10,172**	87.72%
	Car	2	0	4	**6**	66.66%
	Total	**152,879**	**30,726**	**7**	**183,612**	
	Precision	99.1%	29.04%	57.14%		

The confusion matrix for the random forest model is illustrated in table 17. This model not only has high recall for each class but also scores the highest recall for {*Bus*} class, our class of interest with 87.72% recall. The {*Car*} recall has been doubled to reach 66.66%. The {*Bike*} recall has decreased compared to the KNN model but still achieves a high recall rate. The problem with this model is the high number of false positive instances classified as {*Bus*} with 21,803 instances. This high number of false positive instances has lowered the model precision for the class {*Bus*} to be 29.04% which means that only 29.04% out of all the instances predicted by the model as {*Bus*} are actually {*Bus*}. The analysis strategy is to choose the model that lower the number of instances with false negative prediction for our class of interest (high recall) so that the model does not miss any instances which actually labeled as {*Bus*}, after that the instances with false positive predictions (low precision) is filtered with spatial methods using a downstream process. Based on this, the random forest classifier with 87.443% accurate rate that has been trained with balanced data using *Speed* and *Acceleration* features and has n=10 has been chosen to be the best classifier that achieve accepted not only overall accuracy among all classes but also a high recall for the {*Bus*} class which is our class of interest. The model precision for {*Bus*} class is 29.04%. This number is used in the spatial filter process to exclude false positive cases.

4.2.5 Applying the chosen model on MobilityLog segments

In this section, the random forest model that has been chosen from the previous step of analysis is applied to the 3,654 {*motorized*} segments of MobilityLog data. The results of segment classification are that 2,594 segments have been classified as {*Bike*}, 584 segments have been classified as {*Bus*}and 476 segments have been classified as {*Car*}. After adding the 1,234 {*Walk*} segments from section 4.1 then the final result is that 2,594 (53%) segments have been classified as {*Bike*}, 1,234 (25%) segments have been classified as {*Walk*}, 584 (12%) segments have been classified as {*Bus*} and 476 (10%) segments have been classified as {*Car*}. Figure 10 shows the percentage of each transportation mode in MobilityLog data after applying random forest classifiers.

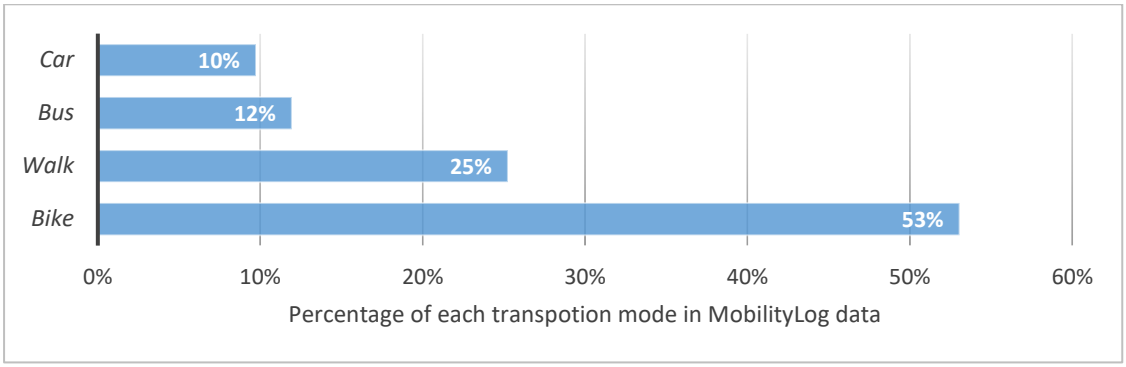

Figure 10. *Distribution of transportation modes in MobilityLog segments*

4.2.6 Mixing machine learning algorithms with geographical analysis

The results of spatial filtering for the false positive MobilityLog {*Bus*} instances are shown in figure 11.

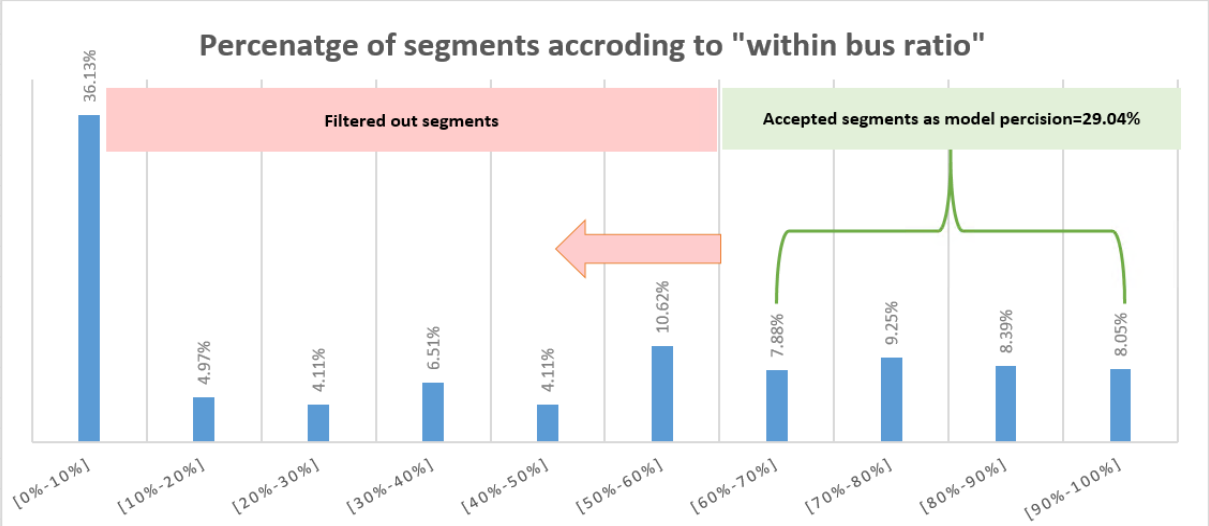

Figure 11. *The filtration process of false positive Bus segments*

According to the random forest model's precision, approximately 30% of the segments are accepted, while the other segments are filtered out. The accepted segments have the highest "*within bus ratio*," which ranges between 70% and 100% (green area), while the segments in the red area have the lower "*within bus ratio*" and have been filtered out.

5 Discussion and conclusion

Smart cities are believed to be the cities of the future. One of the key pillars to build such cities is to design a reliable and sustainable public transportation system that facilitates the mobility of citizens and services while being green and environmentally friendly at the same time. Many classical techniques have been used in the past to plan public transportation systems; however, with the recent revolution in computational power and the availability of the huge amount of data that precisely describe patterns of movements, this method could be updated to leverage the current resources in order to plan more robust and sustainable public transportation systems that satisfy the increasing demand and preserve the environment.

The first step required to conduct this study was obtaining unbiased data that accurately describes people's movements and represents the real usage of different transportation modes in a balanced way. However, acquiring such kind of data was not such a straightforward process. Traditional travel surveys have proved to be an inefficient way to collect the required data as they are time-consuming, inaccurate, and not real-time. The study has shown that data collected using GPS devices and smartphone applications is more convenient, accurate, and easy to collect. However, data usually comes with defects and problems that require huge effort and many iterations to solve. One of the most critical issues is the existence of outliers. Outliers affect the average of measures and can lead to totally different results and observations (Cousineau & Chartier, 2010). Removal of outliers was a crucial part of the data cleaning process during this analysis as the analysis is mainly based on calculating the mean of speed and acceleration for each movement segment and then using these averages to detect different transportation modes through machine learning models. Another important step when dealing with data from different sources like this analysis case is to ensure that all datasets are following the same units and all the timestamps following the same format and time zone. After the cleaning phase, GPS data of each user is divided to many tracks, trips, and segments. The segmentation process models the real movement pattern of users and divides GPS points into chunks that share the same transportation mode. During the analysis, the segmentation process has reduced the data by 99.07% of each original size, and this has significantly lowered the time and computational resources needed to complete the analysis; however, the segmentation process has added more complexity to the data processing steps.

The results also show the ability to use supervised machine learning methods trained by different GPS segment features in order to build models that able to detect different transportation modes in the city of Tartu with a high accuracy rate. Building an accurate machine learning model is a function is many variables such as the cautious choice of training data, the parameter of the model itself, and the feature used during the learning phase. Using an imbalanced dataset during the learning phase results in models with higher accuracy than the same models trained with a balanced training dataset; however, the overall accuracy is not the only measure to consider when testing machine

learning models. Metrics like recall and precision for each individual class are important to measure the sensitivity of the model. Machine learning models that trained with balanced datasets have shown higher recall and precision for all the classes and overall high accuracy. For that reason, only results of models with balanced training datasets are considered even if they show lower overall accuracy than the same models trained with imbalanced datasets. The second important factor in building a robust machine learning model is the selection of the training input features of the movement segments; tested models have shown higher accuracy when using the average acceleration between segments points than the accuracy when using average speed between segments points. However, the same models have shown the highest accuracy in detecting public transport mode when using both average speed and acceleration as training features. The results show that the model that has scored the highest accuracy is a random forest model with a number of trees equals ten and overall accuracy equals 87.443% and 87.72% recall for the public transport detection, followed by a k-nearest neighbours models with k=5 and overall accuracy equals 85.518% and decision tree model ranked last with overall accuracy equals to 84.925%. These results are higher than results achieved in other studies for the same models (Dabiri & Heaslip, 2018). The last thing to consider when building a machine learning model is tuning the model's most important parameters. In the case of the random forest model, the results of the same model with the same setup have been changed according to the number of training trees; the same applies to the KNN model, where accuracy has been varied according to the number of neighbours. Another recommendation to achieve higher accuracy is to train the model with data coming from the same source or even the same users (Shafique & Hato, 2015). This means that models trained with labeled data coming from the MobilityLog application could achieve higher accuracy.

The distribution of the four available transportation mode is the city of Tartu {*Bike, Walk, Bus, Car*} based on the random forest prediction are {53%, 25%, 12%, 10%} respectively, however, this distribution does not imply on the whole city travel pattern and should not be generalized based on the fact that the MobilityLog data is biased towards university students and staff and not all the citizens of Tartu. Finally, the analysis shows the ability to combine the machine learning output with spatial information of the city's bus routes to filter out wrongly predicted bus segments using the model's precision value for the {*Bus*} class. Segments with a larger number of points intersecting with the bus routes are more likely to have {*Bus*} as their transportation mode, while segments with a lower number of points intersecting with the bus routes should be filtered as they are more likely to be wrongly classified by the model. The same framework used in the study can be replicated in other cities in order to detect different transportation modes in general and public transport mode in particular.

Ühistranspordi kasutamise tuvastamine Tartu linnas nutitelefonipõhiste GPS-andmete ja masinõppe meetoditel

Abdelrahman Galal Elnahas

Kokkuvõte

On levinud arvamus, et targad linnad on tuleviku linnad. Selliste linnade loomise üheks põhisambaks on usaldusväärse ja jätkusuutliku ühistranspordisüsteemi kujundamine, mis hõlbustab inimeste ja teenuste liikuvust, olles samal ajal keskkonnahoidlik ja keskkonnasõbralik. Ühistranspordisüsteemide kavandamiseks on varasemalt kasutatud paljusid klassikalisi tehnikaid. Hiljutine arvutusvõimsuse revolutsioon ja suur kättesaadavate andmete hulk, mis täpselt kirjeldavad liikumise mustreid, võimaldavad aga kasutada uuemaid meetodeid. See on vajalik, et suurendada olemasolevat ressurssi ja seeläbi kavandada jõulisemaid ja jätkusuutlikumaid ühistranspordisüsteeme, mis rahuldavad kasvavat nõudlust ja säästavad keskkonda.

Käesoleva magistritöö eesmärgiks on kasutada GPS-toorandmeid ja masinõppe (machine learning – ML) algoritme koos ühistranspordisüsteemi ruumiandmetega nagu bussiliinide asukohad, et luua mudel Tartu linna ühistranspordi kasutamise tuvastamiseks. Eesmärgist lähtuvalt on sõnastatud kolm järgnevat uurimisküsimust.

1. Mis on kõige olulisem(ad) liikumisega seotud tunnus(ed) ühistranspordi kasutamise tuvastamiseks?
2. Milline on suurima täpsusega masinõppe algoritm ühistranspordi kasutamise tuvastamiseks?
3. Kuidas saab masinõppe algoritmide ja klassikalise geograafilise analüüsi kombineerimisel parandada ühistranspordi kasutamise tuvastamise mudelite täpsust?

Treeningandmetena on kasutatud kolme märgistatud GPS-toorandmestikku, et kasutada masinõppe meetodeid nagu k-lähimad naabrid (K-nearest neighbours – KNN), otsustuspuu (Decision Tree – DT) ja otsustusmets (Random Forest – RF). Iga andmestik esindab ühte kolmest Tartu linnas olemas olevast transpordiliigist: buss, auto, jalgratas. Andmestikud puhastati ja segmenteeriti. Segmenteerimine on protsess iga kasutaja GPS-punktide osadeks jagamiseks, kus kõik sama osa punktid on sama transpordiviisiga. 10 puuga otsustusmetsa (RF) mudel osutus kõige paremaks, 87,443 protsendi täpsusega mudeliks. Nii keskmise kiiruse kui ka keskmise kiirenduse kasutamine masinõppemudeli sisendina andsid kõrgema täpsuse võrreldes ainult iga üksiku liikumise tunnuse kasutamisega. Seejärel kasutati otsustusmetsa (RF) mudelit märgistamata MobilityLog mobiilirakenduse andmete jaoks transpordiviiside tuvastamiseks.

Nelja võimaliku Tartu transpordiliigi (jalgratas, jala käimine, buss, auto) jaotus on otsustusmetsa (RF) mudeli tuvastamise põhjal vastavalt 53%, 25%, 12%, 10%. Siiski ei esinda antud jaotus kogu linnas toimuvat liikumist ja seda ei tohiks üldistada, sest MobilityLog'i andmed on kallutatud ülikooli üliõpilaste ja töötajate poole ning ei kajasta kogu Tartu elanikke. Masinõppe meetodid on kombineeritud ka Tartu linna

bussiliinide ruumiandmetega, et filtreerida valesti tuvastatud bussi kasutamise segmendid kasutades mudeli täpsuse väärtust klassi „buss“ jaoks. Lõikudel, kus bussiliinidega ristub suurem arv punkte, on suurema tõenäosusega transpordiliigiks „buss“. Samas kui bussiliinidega ristuvad väiksema punktide arvuga segmendid tuleks filtreerida, kuna need klassifitseeritakse suurema tõenäosusega mudeli poolt valesti.

Acknowledgments

Firstly, I would like to thank Dr.Siiri Silm for her continuous support during the entire thesis. Special thanks for Dr.Amnir Hadachi for his contribution and ideas. Also, I would like to thank the entire family of the institute of geography at university of Tartu as they never hesitated to provide help and support. Finally, I would like to thank my friends for helping and believing in me.

References

(ITU-T), T. I. (2014). *Smart sustainable cities: An analysis of definitions.* The International Telecommunication Union (ITU).

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*.

Bamwesigye, D., & Hlavackova, P. (2019). Analysis of Sustainable Transport for Smart Cities. *Sustainability*.

Behrendt, F. (2016). Why cycling matters for Smart Cities. Internet of Bicycles forIntelligent Transport. *Journal of Transport Geography*, 157-164.

Bricka, S., & Murakami, E. (2012). Advances in travel survey technolog. *Thirteenth International Conference on Travel Behaviour Research.* Toronto, Ontario, Canada.

Chawla, N. (2005). Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 853-867). pringer Science+Business Media, LLC.

Chollet, F. (2017). *Deep Learning with Python.* Manning Publications.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research*.

Dabiri, S., & Heaslip, K. (2018). nferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation Research Part C: Emerging Technologies*, 360-371.

Dijkstra, L., & Poelman, H. (2015). *MEASURING ACCESS TO PUBLIC TRANSPORT IN EUROPEAN CITIES.* European Commission.

Elias, W., & Shiftan, Y. (2012). The influence of individual's risk perception and attitudes on travel behavior. *Transportation Research Part A: Policy and Practice*, 1241-1251.

Fawagreh, K., Gaber , M. M., & Elyan, E. (2014). Random forests: from early developments torecent advancements. *Systems Science & Control Engineering: An Open AccessJournal*, 602-609.

Feng, W., Huang, W., & Ren, J. (2018). Class Imbalance Ensemble Learning Based on the Margin Theory. *Applied Sciences*, 815.

Garbade, M. J. (2018, Auguest 11). *Regression Versus Classification Machine Learning: What's the Difference?* Retrieved from medium: https://medium.com/quick-code/regression-versus-classification-machine-learning-whats-the-difference-345c56dd15f7

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow.* O'Reilly Media, Inc.

Google. (2021). *GTFS*. Retrieved from GTFS: https://developers.google.com/transit/gtfs/reference/

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *Lecture Notes in Computer Science*, 345-359.

Guo, G., Wang, H., Bell, D., & Bi, Y. (2004). KNN Model-Based Approach in Classification.

Hassanat, A. B., Abbadi, M. A., & Alhasanat , A. A. (2014). Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. *International Journal of Computer Science and Information Security*.

Imandous, S. B., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach or Predicting Economic Events:Theoretical Background. *S B Imandoustet al. Int. Journal of Engineering Research and Applications*, 605-610.

Inner City Bus Transportation. (2021, 03 30). Retrieved from Tartu: https://www.tartu.ee/en/inner-city-bus-transportation

Jesmeen, M., Hossen, J., Sayeed, S., Ho, C., Tawsif, K., Rahman, M. A., & Hossain, M. A. (2018). A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics. *Indonesian Journal of Electrical Engineering and Computer Science*, 1234.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Ham, F., Henry Riche, N., . . . Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 271-288.

Knupfer, S. M., Pokotilo, V., & Woetzel, J. (June 2018). *Elements of success: Urban transportation systems of 24 global cities.* McKinsey Center for Future Mobility.

Kohani, A. R. (2017, May 30). *Regression vs Classification*. Retrieved from meduim: https://medium.com/@ali_88273/regression-vs-classification-87c224350d69

Komal, B., Khivsara, S., & Bramhecha, A. (2017). ANDROID APPLICATION USING GPS NAVIGATION.

Komoot. (2021). *Komoot*. Retrieved from Komoot: https://www.komoot.com/

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 249-268.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2005). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 25-36.

L'Heureux, A., Grolinger, K., El Yamany, H., & Capretz, M. (2017). Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*.

Lindenberg, K. (2014). Comparative Analysis of GPS Data. *Undergraduate Journal of Mathematical Modeling: One + Two*, 6.

Liu, Y., Zhou, Y., Wen, S., & Tang, C. (2014). A Strategy on Selecting Performance Metrics for Classifier Evaluation. *International Journal of Mobile Computing and Multimedia Communications*, 20-35.

Luna, J., Cano, A., & Ventura, S. (2016). *A Data Structure to Speed-Up MachineLearning Algorithms on Massive Datasets.*

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). *Big Data: The Next Frontier for Innovation, Comptetition, and Productivity.* McKinsey Global Institute.

Marek, O., Daria, A.-K., & Anna, K. (2020). Sustainable Transport: An Efficient TransportationNetwork—Case Study. *sustainability*.

Medar, R., Rajpurohit, V., & Rashmi, B. (2017). *Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning.*

Minastireanu, E., & Mesnita, G. (2019). An Analysis of the Most Used Machine Learning Algorithms for Online Fraud Detection. *Informatica Economica*, 5-23.

Mitchell, D. (2014). New traffic data sources--An overview. *Bureau of Infrastructure, Transport and Regional Economics, Canberra, ACT, Australia*.

Mo, Y., Zhao, D., Du, J., Syal, M. M., Aziz, A., & Li, H. (2020). utomated staff assignment for building maintenance using natural language processing. *Automation in Construction*.

Nations, U. (2018, May 16). *United Nations*. Retrieved from United Nations: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

Nguyen, T. (2015). Travel survey data: Comparative analysis from different travel survey methods.

Nolan, J. (2009). HISTORY OF GOODS TRANSPORTATION. In T. J. Kim, *TRANSPORTATION ENGINEERING AND PLANNING – Vol. I* (pp. 146,147). Encyclopedia of Life Support Systems.

O'Dea, S. (2019). *Statista*. Retrieved from Statista: https://www.statista.com/statistics/566096/predicted-number-of-smartphone-users-in-estonia/

Oliver, B., Rein, A., Erki, S., & Robert, W. (2017). Extracting Regular Mobility Patterns From Sparse CDR DataWithouta priori Assumptions. *Location Based Service*.

Patel, H., & Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. *International Journal of Computer Sciences and Engineering*, 74-78.

Patil, M. M., & Hiremath, B. N. (2018). A Systematic Study of Data Wrangling. *I.J. Information Technology and Computer Science*, 32-39.

Ren, Q., Cheng, H., & Han, H. (2017). Research on Machine Learning Framework Based on Random Forest Algorithm. *AIP Conference Proceedings.* AIP Publishing.

Saif, M., Maghrour Zefreh, M., & Torok, A. (2018). Public Transport Accessibility: A Literature Review. *Periodica Polytechnica Transportation Engineering*.

Schmöcker, J.-D., Michael, B., & Lam, W. (2003). Importance of public transport. 1-4.

Scotland, T. (2020). *Transport Scotland*. Retrieved from Transport Scotland: https://www.transport.gov.scot/active-travel/developing-an-active-nation/sustainable-travel-and-the-national-transport-strategy/

Shafique, M., & Hato, E. (2015). Formation of Training and Testing Datasets, for Transportation Mode Identification. *Journal of Traffic and Logistics Engineering*, 77-80.

Share, T. S. (2021). *Tartu Smart Bike Share*. Retrieved from Tartu Smart Bike Share: https://ratas.tartu.ee/about

Shen, L., & Stopher, P. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, 316-334.

Simeone, O. (2018). A Very Brief Introduction to Machine LearningWith Applications to Communication Systems. *IEEE Transactions on Cognitive Communications and Networking*, 648-664.

Simon, A., Deo, M. S., Venkatesan, S., & Babu, D. R. (2015). An Overview of Machine Learning and its Applications. *International Journal of Electrical Sciences & Engineering (IJESE)*, 22-24.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 263-286.

Steenbruggen, J., Tranos, E., & Nijkamp, P. (2014). Data from mobile phone operators: A tool for smarter cities? *VUA - Research Memorandum 2014-1*, 1-22.

Stenneth, L. W., Yu, P., & Xu, B. (2011). Transportation Mode Detection using Mobile Phones and GIS Information. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 54-63.

Ting, K. (2017). Confusion Matrix. In C. Sammut, & G. I. Webb, *Encyclopedia of Machine Learning and Data Mining* (p. 260). Boston: Springer.

Vakula, D., & Raviteja, B. (2017). Smart Public Transport for Smart Cities. *International Conference on Intelligent Sustainable Systems.* Telangana: IEEE.

Vieira, D., & Paixao, J. (2018). *Vector Field Neural Networks.*

Wang, Z., He, S., & Leung, Y. (2017). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*.

Wolf, J., Bachman, W., Oliveira, M., Auld, J., Mohammadian, A., Vovsha, P., & Zmud, J. (2014). *Applying GPS Data to Understand Travel Behavior, Volume II: Guidelines.* National Academies of Sciences, Engineering, and Medicine.

Zheng, Y., Liu, L., Wang, L., & Xing, X. (2008). Learning transportation mode from raw GPS data for geographic applications on the Web. *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, 247-256.

Non-exclusive licence to reproduce thesis and make thesis public

I, **Abdelrahman Galal Elnahas** (date of birth: 02.04.1989),

1.herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Detecting public transport mode in the city of Tartu using smartphone-based GPS data and machine learning methods,
supervised by **Dr.Siiri Silm.**

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **24.05.2021**