UNIVERSITY OF TARTU
Institute of Computer Science
Conversion to IT Curriculum

Toomas Gross

Explanatory and Predictive Modelling in the Study of Overweight and Obesity: The Example of Health Behaviour Among Estonian Adult Population

Master's Thesis (15 ECTS)

Supervisor: Rajesh Sharma, PhD

Tartu 2022

Explanatory and Predictive Modelling in the Study of Overweight and Obesity: The Example of Health Behaviour Among Estonian Adult Population

Abstract: The use of machine learning models has become increasingly popular in the study overweight and obesity, complementing research on these topics by means of "traditional" statistical methods. Motivated by this methodological shift as well as the idiosyncrasies of the Estonian context, this thesis has three aims. Building on the data from the 2020 Health Behaviour Among Estonian Adult Population Survey (n = 1,737), it firstly scrutinises the possible associations between being overweight (BMI ≥ 25.0) or obese (BMI ≥ 30.0) and various socio-demographic and behavioural variables through explanatory modelling, using binary logistic regression analysis. Secondly, it compares the performance of various commonly used machine learning algorithms for classification problems when predicting overweight and obesity, respectively. And thirdly, the thesis discusses the advantages and limitations of explanatory and predictive modelling more generally and in the study of overweight and obesity more specifically, entering into dialogue with various other studies that have used these two approaches to the topic.

Keywords: Overweight, obesity, explanatory and predictive modelling, logistic regression analysis, supervised machine learning

CERCS: P160 Statistics, operation research, programming, actuarial mathematics

Seletav ja ennustav modelleerimine ülekaalulisuse ja rasvumise uurimisel Eesti täiskasvanud rahvastiku tervisekäitumise näitel

Lühikokkuvõte: Masinõppe mudelite kasutamine on üha populaarsem lähenemine ülekaalulisuse ja rasvumise uurimisel, olles täienduseks "traditsiooniliste" statistiliste meetodite rakendamisele nende teemade analüüsis. Ajendatuna taolisest metodoloogilisest "nihkest" ning samuti Eesti konteksti eripäradest, on käesoleval magistritööl kolm eesmärki. Tuginedes kõige hiljutisematele, 2020. aasta Eesti täiskasvanud rahvastiku tervisekäitumise uuringu andmetele (n = 1737), analüüsitakse magistritöös esiteks binaarse logistilise regressiooni mudelite abil ülekaalulisuse (KMI ≥ 25,0) ja rasvumise (KMI ≥ 30,0) võimalikku seost erinevate sotsio-demograafiliste ja käitumuslike teguritega. Teiseks võrreldakse magistritöös kuue erineva klassifitseerimisprobleemide lahendamisel laialt kasutatud masinõppe mudeli võimet ülekaalulisust ja rasvumist treeningandmete põhjal ennustada. Kolmandaks käsitletakse töös kahe eelpool mainitud lähenemise – seletava ja ennustava modelleerimise eeliseid ja puudusi üldisemalt ning ülekaalulisuse ja rasvumise uurimisel konkreetsemalt, astudes seejuures dialoogi ka teiste uurimustega, mis neid lähenemisi antud valdkonnas on kasutanud.

Võtmesõnad: Ülekaalulisus, rasvumine, seletav ja ennustav modelleerimine, logistiline regressioonanalüüs, juhendatud masinõpe

CERCS: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Table of contents

1 Introduction 4
2 Theoretical background and research context 8
2.1 Body mass index and weight categories 8
2.2 Two opposing “modelling cultures” 8
2.3 Explanatory modelling in the study of overweight and obesity 12
2.4 Predictive modelling in the study of overweight and obesity 17
3 Data and methods 21
3.1 Description of the original dataset 22
3.2 Data pre-processing 23
3.3 Methods of data analysis 27
3.3.1 Exploratory data analysis 27
3.3.2 Explanatory modelling 28
3.3.3 Predictive modelling 28
4 Results 34
4.1 Descriptive statistics 34
4.2 Explanatory modelling 39
4.3 Predictive modelling 42
5 Discussion 47
5.1 Explanatory modelling 47
5.2 Predictive modelling 48
5.3 Reconciling the two “modelling cultures” 50
6 Conclusions 52
7 References 54
8 Appendices 60
8.1 Glossary 60
8.2 License 63

1 Introduction

The world's population with excess body weight has increased considerably and steadily in recent decades, in both absolute and relative terms. Being overweight (Body mass index or BMI = 25.0–29.9 kg/m^2) or obese (BMI ≥ 30.0 kg/m^2) have become major "lifestyle diseases" in most countries (Chatterjee et al. 2020, 1). The global nature of the so-called "obesity epidemic" was first formally recognised by the World Health Organization (WHO) in 1997 (Caballero 2007, 1). Although there exist significant regional and demographic variations in the manifestations of this "epidemic" (eg. Colmenarejo 2020; Ezzatti 2017), most studies have argued that the phenomenon now characterises almost all societies and socio-demographic groups (Nuttall 2015, 121). According to some estimates, since 1975 the worldwide obesity has nearly tripled (World Health Organization 2021). In 2016, roughly 1.9 billion adults (aged 18 years and older) were overweight and among these over 650 million were obese, based on the most recent official data available by the World Health Organization. In relative terms, these figures amounted to 39 and 13 percent of the world's population, respectively. The World Health Organization also estimates that over 340 million children and adolescents (aged 5 and above) were overweight or obese in 2016.[1] Turning the gaze more specifically on Europe – based on the most recent European Health Interview Survey (EHIS, third wave), more than half (53 percent) of the adult population in Europe was overweight or obese in 2019 (Eurostat 2021).[2] France and Italy were the only EU countries where *less* than half of the adult population was overweight or obese, according to this survey.

Obesity in particular but excess body weight also more generally, constitute major public health concerns. Both are associated, first are foremost, with certain types of cancer, type 2 diabetes, obstructive pulmonary, various cardiovascular, and many other diseases (Cañas Cervantes and Martinez Palacio 2020; Chatterjee et al. 2020; Cheng et al. 2021; Csige 2018; Delnevo et al. 2021; Thamrin et al. 2021). The World Health Organization estimates that by 2030 up to 30 percent of all global deaths could be caused, directly or indirectly, by lifestyle diseases, especially obesity (Chatterjee et al. 2020, 1). Understanding the causes of excess body weight has thus very practical relevance in terms of health care and behavioural interventions, and for planning actions that contribute to raising awareness about its risk factors and mitigating measures.

Estonia is no exception to these world- and Europe-wide trends and constitutes a particularly interesting context for studying the topics of overweight and obesity. According to the most recent official estimates based on the weighted results of the Health Behaviour Among Estonian Adult Population survey in 2020, 51.6 percent of Estonian adult population aged 16–64 are either overweight or obese (BMI ≥ 25.0 kg/m^2) and 20.5 are obese (BMI ≥ 30.0 kg/m^2) (Reile and Veideman 2021, 91). With these figures, Estonia is in the forefront of the European countries in terms of overweight and obesity prevalence among its population, especially in some socio-demographic groups. This has been the case for quite some time. A comparative study from already more than a decade ago, for example, demonstrated that Estonia and Latvia had the highest level of obesity among men in Europe (Webber et al. 2010). A more recent study listed Estonia as the second most obese country in the European

[1] Afshin et al. (2017) suggest, based on the BMI data from 195 countries, that in 1990-2015 childhood obesity levels in many countries increased actually faster than those of adults.

[2] The survey covers the member states of the European Union and the countries of the European Economic Area.

Union after Slovenia (Marques et al. 2017). The above-mentioned European Health Interview Survey, in turn, put Estonia among the top seven most obese countries in the European Union (Eurostat 2021).

Since 2010, based on the results of the Health Behaviour Among Estonian Adult Population surveys, the proportion of the overweight *and* obese population (BMI ≥ 25.0 kg/m^2) among Estonian adults has remained fairly stable, while the share of obese population (BMI ≥ 30.0 kg/m^2) has slightly increased, especially since 2018 (Reile, Tekkel, and Veideman 2019; Reile and Veideman 2021; Tekkel and Veideman 2011, 2013, 2015, 2017). Interestingly, this coincides with various more positive aggregate trends in health and physical behaviour during this era. For example, the share of Estonian adults exercising at least once a week has considerably increased since 2010. The percentage of regular smokers among Estonian adults indicates a clear downward trend. The share of the adult population consuming alcohol at least weekly, although remaining fairly stable throughout the decade, appears also to have dropped towards the end of it. These parallel trends are summed up on Figure 1 below, and their counterintuitive simultaneousness allows one to assume that the prevalence of overweight and obesity in Estonia is a complex phenomenon that deserves to be approached and scrutinised from different methodological perspectives.

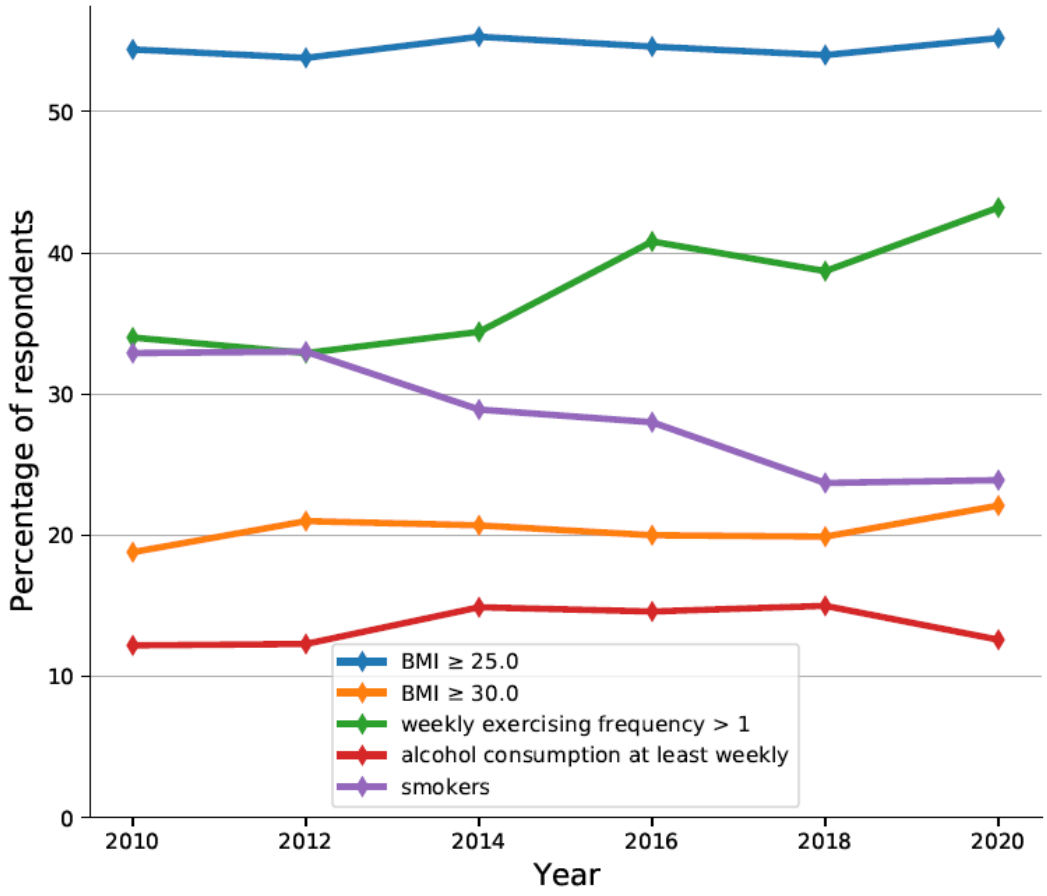


Figure 1
Trends of Overweight and Obesity Prevalence and Selected Behavioural Features in Estonia (Health Behaviour Among Estonian Adult Population Surveys, 2010–2020)

Explanatory research on overweight and obesity as "lifestyle diseases" has generally focused on their risk factors. These studies have by now a long history, the body of such research is extensive and it continues to expand, scrutinising excess body weight as a complex multifactor phenomenon with many causes. More recently, however, explanatory studies have been increasingly complemented by research that strives to predict overweight and obesity condition from available data using the computational approaches of machine learning models (Cañas Cervantes and Martinez Palacio 2020, 5–14; Chatterjee et al. 2020, 1; Colmenarejo 2020; Thamrin et al. 2021, 1).

Motivated by such shift in research trends from explanatory towards more predictive approaches as well as the idiosyncrasies of the Estonian context, this thesis, building on the most recent Health Behaviour Among Estonian Adult Population survey data from 2020 (n = 1,737), has threefold research aims:

1) To scrutinise the possible associations between being overweight (BMI $\geq$ 25.0 kg/m^2) or obese (BMI $\geq$ 30.0 kg/m^2) and various socio-demographic and behavioural variables by means of explanatory modelling, using binary logistic regression analysis.
2) To compare the performance of six commonly used machine learning algorithms for classification problems when predicting overweight and obesity status, respectively. To the best of the author's knowledge, machine learning has not been used for these purposes in case of Estonian data.
3) To discuss the advantages and limitations of explanatory and predictive modelling both more generally and in the study of overweight and obesity more specifically, while engaging with various other studies that have used these approaches to the topic.

Three conceptual and terminological clarifications are in order here. Firstly, the distinction between "explanatory" and "predictive modelling" – the terms borrowed from Shmueli (2010) – is not clear-cut, as will be highlighted below in Chapter 2. Also, various other terms have been used when outlining different "modelling cultures" – Breiman (2001), for example, distinguishes between "data modelling" and "algorithmic modelling." Simplifying and generalising crudely, the former stands for "traditional" or "classical statistics" and the latter for "machine learning" (Molina and Garip 2019, 29). Secondly, (binary) logistic regression is purposefully employed in this thesis for both explanatory and predictive purposes, and its "double identity" as a "traditional statistical method" (Cheng et al. 2021, 1) that has been appropriated by machine learning is further addressed in section 2.2. And thirdly, throughout this thesis, the category "overweight," unless stipulated otherwise, refers to "BMI $\geq$ 25.0 kg/m^2." This means that the term comprises both overweight (BMI = 25.0–29.9 kg/m^2) and obese categories (BMI $\geq$ 30.0 kg/m^2) as they are defined by the World Health Organization (2000). But since the category "BMI = 25.0–29.9 kg/m^2" on its own is not a focus of analysis in this study, such terminological simplification is justified as it enables to avoid using more cumbersome constructs like "overweight/obese," "overweight and obese," or "at least overweight" when referring to the category "BMI $\geq$ 25.0 kg/m^2".

The thesis is structured as follows: after this introduction, Chapter 2 gives a brief overview of BMI as a metric, the distinction between "explanatory" and "predictive modelling," and research on overweight and obesity, comparatively and in Estonia, from these two perspectives. Chapter 3 presents data and methods, introducing the original dataset, outlining data pre-processing phases and rationales, as well as the statistical approaches, the machine

learning algorithms compared in this study, model training and testing principles, and the employed evaluation metrics. The results of data analysis are presented in Chapter 4. The discussion in Chapter 5 addresses how the results of the logistic regression analysis could be interpreted and how its predictive power compares with that of other machine learning algorithms, also entering into dialogue with various other studies that have endeavoured to explain and predict overweight and obesity. The concluding chapter 6 offers a concise wrap-up of the results and the main points of the study, also highlighting its limitations and suggesting ideas for further research. The appendices of the thesis include a glossary of core terms and concepts, and the license information.

2 Theoretical background and research context

This chapter provides an overview of the conceptual and theoretical background of the study and introduces some of the previous research on overweight and obesity from both explanatory and predictive perspectives.

2.1 Body mass index and weight categories

The most common although not the only metric used to group individuals into weight categories is the Body mass index (BMI), calculated as the person's body weight in kilograms divided by the person's squared standing height in meters and measured in the units of kg/m^2. Based on the single-cut definitions of this metric, the World Health Organization (2000) distinguishes between four main BMI categories:

- underweight (BMI < 18.5 kg/m^2)
- normal weight (BMI = 18.5–24.9 kg/m^2)
- overweight (BMI = 25.0–29.9 kg/m^2)
- obese (BMI ≥ 30.0 kg/m^2)

Obesity is sometimes further divided into three sub-categories: obesity class I (BMI = 30.0–34.9 kg/m^2), obesity class II (BMI = 35.0–39.9 kg/m^2), and obesity class III (BMI ≥ 40.0 kg/m^2) (Chatterjee et al. 2020; Jindal et al. 2018).

It is worth noting that despite the wide acceptance of BMI as a predictor of several health issues and its ubiquitous use in public health policies, the metric has also been criticised – for example, for being a poor indicator of the percentage of body fat and for being highly dependent on age as well as gender that can significantly affect the interpretation of BMI values, particularly with regard to health risks (Chatterjee et al. 2020, 2; Jindal et al. 2018, 356; Nuttall 2015, 124). In case of children and adolescents in particular, the BMI categories can be misleading and often an age- and gender-based distribution of the population BMI is used as a point of reference instead (Colmenarejo 2020, 14). Another obstacle to an unequivocal interpretation of the BMI values is that standing body height is considerably influenced by the length of legs while body weight is influenced by both muscles and fat (Nyholm 2007, 197).

Because of these limitations, some authors have called for changing the BMI classification system (Nuttall 2015), or for using alternative metrics such as the waist-hip ratio, for example (Chatterjee et al. 2020; Jindal et al. 2018; Wanner et al. 2016). The limitations of the BMI as a metric need to be acknowledged also in the context of this study although the fact that its focus is on adult population and that the BMI is used here as a dependent variable and not as a predictor of health issues or body composition, largely mitigate the above-described deficiencies.

2.2 Two opposing "modelling cultures"

This section gives a brief but critical overview of the distinction between explanatory and predictive or what Breiman (2001) in his seminal article called "data modelling" and

"algorithmic modelling cultures" by way of introduction to the next two sections that outline how these approaches have been employed in various studies of overweight and obesity.

The most essential difference between these two modelling cultures is often boiled down to the differences between their core aims – those of explanation and prediction (Breiman 2001; Buskirk et al. 2018; Bzdok et al. 2018; Shmueli 2010). In reality, however, the distinction between the two is not clear-cut (Christodoulou et al. 2019; Shmueli 2010;), and scholarly approaches to whether such juxtaposition is justified may differ. While Breiman (2001) writes about two clearly different "cultures," other authors have defined machine and deep learning as fields "at the intersection of statistics and computer science" (Molina and Garip 2019) and "at the interface between statistics and artificial intelligence" (Colmenarejo 2020, 3), or conceptualised these as nothing more but simply "natural extensions to conventional statistical methods" (eg. Chatterjee et al. 2020, 6).

In explanatory or data modelling, as Breiman (2001, 203) calls the approach in his article, the aim is to extract information from the data about the underlying mechanism producing it. This type of modelling is generally performed for testing causal hypotheses, hence the prominent role of theory in it. Such an endeavour requires interpretable statistical models that are easily linked to the underlying theoretical model (*ibid*., 298). As Buskirk et al. (2018, 3) add, in social sciences in particular, a relevant underlying theoretical model posits a functional, generally causal relationship between the constructs and an outcome of interest. These constructs are operationalised into variables used in the explanatory model and the model thus constructed is aimed to be parsimonious, while having maximal explanatory capacity (*ibid*.). The inclusion of important predictors in the final model is often quantified in terms of effect size measures, confidence intervals, or p-values for estimated coefficients, and the outcome is often a relatively simple and comprehensible picture of the relationship between input variables and responses (*ibid*.). From among many examples of this modelling culture, Breiman (2001, 203) explicitly singles out the classificatory tool of logistic regression that produces a linear combination of the variables with weights that indicate variable importance, thus offering an intuitively good sense of how independent variables affect the dependent one. DeGregori et al. (2017, 671) similarly argue that the strength of various types of regression models lies in their well-understood theoretical and computational background.

High interpretability explains why statisticians, according to Breiman (2001, 204), seem to be stuck with various types of regression analysis methods, despite their significant limitations. For example, these models generally assume normally distributed data (DeGregori et al. 2017, 671). As Breiman (2001, 204) provocatively suggests, "nobody really believes that multivariate data is multivariate normal […], and the a priori assumption that nature would generate the data through a parametric model selected by the statistician can result in questionable conclusions that cannot be substantiated by appeal to goodness-of-fit tests and residual analysis." Also, relying on regression approaches requires a good knowledge of both data properties and model capabilities in order to be able to successfully implement these models (*ibid*.). This, however, can be difficult if not impossible since data – medical and behavioural data in particular – are often complex and multidimensional. In consequence, imposing simple parametric models on complex data can result in a loss of accuracy or worse – explanatory models might lead to irrelevant theory and questionable scientific conclusions, as Breiman (*ibid*.) maintains. With obesity and overweight studies specifically in mind, Thamrin et al. (2021, 1) echo Breiman's criticism suggesting that the main problem with the

"traditional" regression approach is that it limits analysis to a small number of predictors and imposes assumptions of independence and linearity.

Predictive modelling based on machine learning (with deep learning as its outgrowth), or what Breiman calls "algorithmic modelling culture," has emerged in recent decades as an increasingly popular and often deemed as preferable and more feasible alternative to explanatory models in case of studying large, complex, and multidimensional data. The approach has been particularly prominent in the fields of natural language processing and bioinformatics, speech, handwriting and image recognition, prediction of financial markets, etc. (Colmenarejo 2020, 3). Not surprisingly, healthcare data (Chatterjee et al. 2020; Doupe et al. 2019; Vaishya et al. 2020), medical and clinical data (Christodoulou et al. 2019; DeGregory et al. 2018; Ngiam and Khor 2019), and physical activity data (Ahmadi et al. 2020; Ahmed and Loutfi 2013; Willetts et al. 2018; Zhou et al. 2019) – admittedly also complex and multidimensional – are now likewise studied by means of machine learning. The same increasingly applies to the topics of overweight and obesity, as demonstrated below in section 2.4.[3]

While explanatory statistical models are based on various theoretical assumptions, build on the researcher's prior knowledge about the domain, often make inferences based on a sample, and generally test predefined hypotheses, machine learning strives to find generalisable predictive patterns from data by employing highly flexible nonparametric methods (Bzdok et al. 2018, 23; Molina and Garip 2019, 40). The priority in case of such modelling is to generate accurate predictions of new observations and in this process the underlying statistical model is often unknown to the researcher. The prime value of these models resides namely in their predictive accuracy and not necessarily in the interpretability of the model. As Breiman (2001, 204) eloquently puts it, the approach builds on the acknowledgement that "nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable."

At a high level, according to Doupe (2019, 809), applying a machine learning algorithm involves at least five steps: data preparation, estimator family selection, estimator parameter learning, estimator regularisation, and estimator evaluation. The techniques of machine learning are generally divided into two broad categories although a third category is sometimes also distinguished (Buskirk et al. 2018, 1):

- *Supervised learning*. To put it simply, the goal of supervised learning is to predict a dependent variable (also referred to as output, target, class, or label) as a function of one or many independent variables (also referred to as inputs, features, or attributes) (Buskirk et al. 2018, 2). A supervised machine learning model is trained based on a set of training samples paired with the corresponding labels (Ahmed and Loutfi 2013, 1). The model is then tested by assigning class labels on a set of testing samples where the label is unknown, and the prediction results are evaluated by means of different metrics (*ibid*.). The models are constructed with the primary purpose of predicting either continuous or categorical outcomes (Buskirk et al. 2018, 4). Prediction of continuous numeric (or quantitative) output variables is usually referred

[3] Extensive overviews of a vast array of topics that have been approached by employing machine learning in social sciences and especially in survey research can be found in Buskirk et al. (2018), Kern et al. (2019), and Molina and Garip (2019).

to as a "regression problem," whereas prediction of categorical (or qualitative) output variables is referred to as a "classification problem."

- *Unsupervised learning.* In unsupervised learning, unlike in supervised learning, no prespecified dependent variables exist, and the methods focus on detecting patterns among all variables of interest in a dataset with a more general aim to describe or characterise the data, for example by dividing data into different clusters (DeGregory et al. 2018, 669).

- *Reinforcement learning* is sometimes singled out as the third category of machine learning, although in essence it is an intermediate mix of the previous two, where training data are assumed to provide only an indication as to whether an action is correct or not rather than determine the correct output for a given input, and correct actions are rewarded while incorrect actions are punished (Jordan and Mitchell 2015, 258).

In this thesis only supervised machine learning models for classification problems will be used.

When juxtaposing "data modelling" and "algorithmic modelling cultures" in his seminal article, Breiman (2001, 214) was adamant about his preference. Through various examples he demonstrated that the latter offers a far better accuracy, even though this might come at the expense of the researcher's understanding of the underlying data mechanism. Many have followed his suit when stressing the advantages of machine learning over "traditional statistics" – not only because of its better predictive capacity but also because of the ease of its application and the relative robustness of algorithmic models (Bzdok et al. 2018, 233; Cheng et al. 2021; Christodoulou et al. 2019, 13; Colmenarejo 2020). The data-driven nature of machine learning often also means that it does not require very specific prior knowledge of the study domain and it can be easily adapted to complex non-linear interrelations between the outcome and its predictors (Doupe et al. 2019, 814: Kern et al. 2019; Molina and Garip 2019, 37). This is particularly advantageous in case of the so-called "wide data" and big datasets, although more ardent proponents of machine learning have argued that the techniques developed for big data can outperform standard statistics, especially regression models, also when dealing with relatively small datasets (eg. Hindman 2015).

That said, nearly all the mentioned qualities of predictive modelling by means of machine learning can simultaneously be considered its weaknesses (Doupe et al., 2019, 814). Prioritising predictive accuracy, most machine learning and especially deep learning algorithms can be viewed as "black-box" type models, the operating principles of which are difficult if not impossible to interpret. Their application may require considerable investment of computational resources. Moreover, machine learning methods may have deceptively high accuracy but predict the wrong outcome (*ibid*., 815). Another realm of criticism is concerned with the fact that supervised machine learning models in particular can be prone to perpetuate certain social inequalities if applied in practical policy-making contexts (Molina and Garip 2019, 33). Also, as Radford and Joseph (2020) aptly demonstrate, the fact that machine learning is often essentially "theory-free" and does not require the same amount of domain-specific knowledge as classical statistics does, can be a big caveat, at least for meaningful research in social sciences. As they (*ibid*., 10) argue, privileging machine learning models that perform well over models that are founded in a deeper understanding of the society or topic under study puts one in danger of advancing only computer science rather than both

computer and social sciences (or any other discipline). Moreover, it may force the use of machine learning when analysing social data toward "pseudoscience," where algorithms are employed to make potentially discriminatory decisions or social scientific claims that can be baseless (*ibid*.). Instead, as Radford and Joseph (*ibid*.) conclude, (social) theory should remain an integral part of "each step of the machine learning pipeline."

By way of conclusion to this subsection, it is worth asking whether such crude juxtaposition of "data modelling" and "algorithmic modelling cultures," to use Breiman's (2001) terminology, is even justified. While some methods used in these "cultures" fall unequivocally into one or the other domain, many are used in both, such as the bootstrap method that can be used for making statistical inferences, but it also serves as the basis for ensemble methods in machine learning, such as the random forest algorithm (Bzdok et al. 2018, 234). Different types of regression models are also good cases in point. To bring an example of logistic regression modelling employed in this thesis – many authors treat it as a "traditional statistical approach" due to the way features included in the model are selected and adjusted for, and how effect sizes or odds ratios are calculated (eg. Colmenarejo 2020; Cheng et al. 2021). The analysis of the output of logistic regression model essentially serves explanatory and inferential purposes. For instance, using an example based on the data in this thesis – one might try to assess whether and how the education level is associated with having excess body weight while the model is adjusted for other relevant covariates, and whether the putative association is statistically significant. But since a logistic regression model can also be used to predict the log-odds of a binary variable (eg. being obese or not), it can be trained as a machine learning algorithm, and its performance and predictive power can be evaluated by different metrics just like in case of other (supervised) machine learning algorithms (Colmenarejo 2020, 24; Zhang et al. 2009). The "double identity" of logistic (and other) regression models thus further underlines the certain arbitrariness that is inherent to the juxtaposition of the two "modelling cultures" outlined above and the fact that in essence they form a continuum.

2.3 Explanatory modelling in the study of overweight and obesity

Before moving on to presenting and analysing the data that this thesis builds on, a brief and selective overview of research on overweight and obesity, conducted from the perspectives of both explanatory and predictive modelling, is in order since one of the aims of the thesis is to eventually enter in dialogue with some of these studies.

The body of research employing explanatory modelling when studying the role of various factors affecting overweight and obesity is by now immense. Thamrin et al. (2021, 2) divide these factors (and respective research foci) into three groups: socio-demographic, lifestyle, and genetic. Research on socio-demographic factors has mostly focused on gender, age, education, income, marital status, urban-rural residence, race, and ethnicity (eg. Clarke et al. 2008; Seo and Li 2009; Tekkel, Veideman, and Rahu 2010). Research on lifestyle factors has generally paid attention to such behavioural aspects as unhealthy diet and habits (smoking, consumption of alcohol, and drug use, for example), as well as low levels of physical activity (eg. Besson et al. 2009; Fogelholm and Kukkonen-Harjula 2000; Johnson, Kuh, and Hardy 2015; Townshend and Lake 2017). The genetic factors highlighted in explanatory research on overweight and obesity have mainly included various hereditary aspects of the phenomena (eg. Katus et al. 2020). In reality, however, most explanatory studies have acknowledged the multifactor essence of excess body weight and seen its causes to be some combination of

factors from all three mentioned groups (eg. Afshin et al. 2017; Bixby 2019; Courtemanche 2016; Kirby 2012; Paeratakul et al. 2002; Reile et al. 2020).

Caution is needed when making generalisations about the outcomes of this vast body of research, as the studies have been conducted in very different socio-demographic contexts and with a focus on different independent variables. Moreover, the potential factors related to overweight and obesity vary considerably depending on whether the focus is on children, adolescents, or adults (eg. Chatterjee et al. 2020, 2; Ezzatti 2017). Yet there is an important methodological similarity to many studies that have used "traditional statistical modelling" when *explaining* overweight and obesity – it has been common to employ regression analysis of some sort when doing this (Colmenarejo 2020, 24).

There is no space here to provide a comprehensive overview of explanatory research on overweight and obesity, but some versions of this can be found in Colmenarejo (2020) and Cañas and Martinez (2020). Only a few illustrative cases – from both Estonian and other contexts – are singled out below, as these studies will be returned to in the discussion chapter of the thesis.

- In an earlier study based on cross-sectional data from different Health Behaviour Among Estonian Adult Population surveys (1990–2004), Tekkel, Veideman, and Rahu (2010) scrutinised the changes in the trends of obesity in Estonia. Using logistic regression and focusing on four socio-demographic variables (education, residence, ethnicity, and income) and the use of outpatient healthcare services by obese individuals as independent variables, the study concluded that the changes in obesity prevalence in Estonia over this period correlated well with economic changes, although from the studied variables, only age and education (among women) had a strong relationship with long-term changes in obesity levels. The use of outpatient medical care by obese individuals and those with "normal" BMI differed only slightly.

- In another study based on Health Behaviour Among Estonian Adult Population survey data, Reile and Leinsalu (2019) analysed the trends of and socio-demographic factors associated with self-reported weight-reducing behaviours such as dietary habits and physical activity among individuals with excess body weight during the period of 2006–2016. Building on descriptive statistics and using multivariate logistic regression, the study concluded that less than half of overweight or obese individuals reported having changed their eating habits or physical activity during the previous twelve months and that improvements in weight-related behaviour were more common among younger adults, among women, and among respondents who did not engage in risky health behaviour.

- In yet another study based on Health Behaviour Among Estonian Adult Population survey data, Reile et al. (2020) analysed age, period, and cohort effects on the mean BMI and obesity over the period of 1996–2018. The authors used hierarchical age-period-cohort analysis with cross-classified random effects modelling, performed separately for men and women. The study established a curvilinear association between age and mean BMI for men, and a linear association of that for women, concluding that population-level BMI changes in Estonia during the studied period were mostly driven by period rather than cohort-specific changes.

- In a study focusing on all three Baltic countries, Oja et al. (2020) assessed the time trends in the adolescents' physical activity, dietary habits, and BMI group (overweight or obese), building on the Health Behaviour in School-aged Children (HBSC) study data for the period of 2006-18. Using logistic regression, the study concluded that the patterns of health behaviour of adolescents in Estonia, Latvia, and Lithuania since 2006 had been fairly similar, although some differences did exist. In Estonia in particular, the adolescents' physical activity tended to be relatively lower and overweight and obesity rates higher than in Latvia and Lithuania.

- To take an example beyond the Estonian and Baltic context: Clarke et al. (2008) aimed in their research to diverge from the frequently used cross-sectional or single cohort study design when studying obesity and overweight by using longitudinal panel data with repeated measures from the nationally representative Monitoring the Future Study (1986–2004) in the United States in order to examine social disparities in the trajectories of BMI over adulthood (ages 18–45). Using growth curve models, the authors analysed these trajectories by gender, race/ethnicity, and lifetime socio-economic position (measured by parents' and respondent's education). The BMI trajectories exhibited a curvilinear rate of change, but a strong period effect was detected – weight gain was more rapid for more recent cohorts. The results of the study highlighted the importance of social status and socioeconomic resources for maintaining optimal weight, but also indicated that even the population in advantaged social positions had experienced an increase of BMI in recent years.

- Maher et al. (2013) examined, cross-sectionally, the combined influence of moderate-to-vigorous physical activity (MVPA) and sedentary behaviour on obesity among 5,083 adults in the United States who had responded to the National Health and Nutrition Examination Survey (2003–5). The independent associations between MVPA, television and total sedentary time, and obesity were examined using logistic regression, and seven socio-demographic variables (age, ethnicity, education level, household income, smoking status, mean daily energy and alcohol intake) were included as covariates in the regression models. As the study demonstrated, the MVPA was consistently inversely associated with obesity, regardless of the nature or the amount of sedentary behaviour. There existed inconsistent positive associations between television time and risk of obesity in men, but not between total sedentary time and risk of obesity in either men or women. As the study concluded, obesity was more strongly related to the MVPA than either television time or total sedentary time. Small differences in daily MVPA, in turn, were associated with relatively large differences in risk of obesity.

- Seo and Li (2009) explored the dose-response effects of leisure-time physical activity (LPTA) on obesity among 12,227 adults aged 20–64 years, drawn from eight years (1999–2006) of the National Health and Nutrition Examination Survey in the United States, employing logistic regression and using leisure-time physical activity as the primary independent variable. As a potential confounder, a measure of occupational physical activity was included in the analysis. Logistic regression models were adjusted for age, gender, and race/ethnicity. The study concluded that there was a crude graded inverse dose-response relationship between the total volume of LTPA and obesity among the US adult women, but not among men. Gender and racial/ethnic differences existed in the relationship of accumulated LTPA with obesity.

- Stamatakis et al. (2008) investigated the relationships of physical activity types and sedentary behaviour with BMI and waist circumference (WC), studying a sample of 6,215 adults aged 16 years and older in Scotland. The dependent variables used in their study were BMI-defined obesity and WC-defined obesity. In their backward stepwise multiple logistic regression models, self-reported average daily television and other screen-based entertainment and moderate to vigorous intensity time were the two main independent variables, and the models were adjusted for eight different confounding variables (sex, age, socio-economic status, self-reported health status, consumption of soft drinks, energy-dense snacks and alcohol, occupational physical activity). The study concluded that physical activity and sedentary behaviour were independently related to both ways of defining obesity.

- The study by Wanner et al. (2016) had twofold aims: firstly, to scrutinise cross-sectional associations between domain-specific physical activity, sedentary time, and different objective measures of overweight and obesity; and secondly, to examine the longitudinal associations between respondents' patterns of change in physical activity, and overweight and obesity ten years later. The research was based on the first (2002/3) and the second (2010/11) follow-up of a Swiss cohort study (SAPALDIA). Multivariate logistic regression models were used for all analyses with binary overweight or obesity as outcome variables (either overweight versus normal weight or with obesity versus normal weight), with physical activity as the primary independent variable and ten socio-demographic, health- and behaviour-related variables as potential confounders. The study results revealed associations, to a varying degree, between physical activity, and overweight and obesity cross-sectionally and longitudinally.

The core methodological aspects of these nine studies are summarised in Table 1 on the next page.

Table 1

Examples of Studies on Overweight and Obesity Using Explanatory Modelling

Study (year)	Country	Age group	Sample size	Dependent variables	Independent variables	Method(s)	Type of study
Tekkel et al. (2010)	Estonia	16-64	11,774	Obesity, use of health services by the obese population	4	LoR	CS, L (1990-2004)
Reile and Leinsalu (2020)	Estonia	20-64	7929	Eating habits and physical activity among overweight/obese population	12	LiR, LoR	CS, L (2006-2016)
Reile et al. (2020)	Estonia	16-64	27,845	Mean BMI, obesity	3	LoR	APC (1996-2018)
Oja et al. (2020)	Estonia, Latvia, Lithuania	11-15	56,340	Physical activity, dietary habits, overweight/obesity	13	LoR	CS, L (2006-2018)
Clarke et al. (2008)	US	18-45	10,956	BMI gain over time	3	LiR	L (1986-2004)
Maher et al. (2013)	US	20-	5,083	Obesity	9	LoR	CS (2003-2005)
Seo and Li (2009)	US	20-64	12,227	Obesity, use of health services by the obese population	5	LoR	CS, L (1999-2006)
Stamatakis et al. (2008)	UK/Scotland	16-	6,215	Obesity defined by BMI and WC	10	LoR	CS (2003)
Wanner et al. (2016)	Switzerland	18-60	13,968	Obesity, WC, WHR, BIA body fat	11	LoR	CS, L (2002-2011)

Note. Abbreviations: APC: Age-period-cohort study; BIA: Bioelectrical impedance analysis; CS: Cross-sectional study; L: Longitudinal study; LiR: Linear regression; LoR: Logistic regression; ND: Not described; WC: Waist circumference; WHR: Waist-to-height ratio

2.4 Predictive modelling in the study of overweight and obesity

As already mentioned above, machine learning has emerged in recent decades as an increasingly common alternative to the more "traditional" statistical approaches when studying overweight and obesity (Thamrin et al. 2021, 1). The body of research aiming at predicting overweight and obesity status and comparing the performance of different algorithms is by now vast and heterogeneous, and just like in case of explanatory research outlined in the previous section, it is difficult to offer a concise overview of these studies. To generalise crudely, both supervised and unsupervised machine learning models have been employed, although the former tend to dominate in the studies. Solving classification problems has been more common than solving regression problems, i.e. these studies have more often focused on predicting the BMI group than the exact BMI value. Among the former, most studies have focused on predicting obesity and only some (eg. Cheng et al. 2021) have separately also focused on the overweight category (i.e. BMI = 25.0-29.9 kg/m^2). While some studies have employed only one single method (eg. Jindal et al. 2018; Kim et al. 2019; Pang et al. 2019; Selya and Anshutz 2018), others have aimed to evaluate and compare the predictive capacity of a smaller (eg. Dunstan et al. 2020; Zheng and Ruggiero 2017) or a bigger set (eg. Cheng et al. 2021; DeGregory et al.. 2018) of algorithms. It is difficult to single out the most popular algorithms employed, although among the recurring ones are the algorithms also compared in this thesis: support vector machine, different versions of decision trees, k-nearest neighbours, random forest, naïve Bayes, as well as logistic regression. The "double identity" of the latter – addressed above and made practical use of in this thesis – is also evident when scrutinising its use in other studies. Although most research accounts have employed logistic regression exclusively as a machine learning model (eg. Rios-Julian et al. 2017; Thamrin et al. 2021; Zhang et al. 2009; Zheng and Ruggiero 2017), some studies (eg. Colmenarejo 2020; Cheng et al. 2021) have approached it as an example of statistical modelling instead.

The results of the research using and comparing predictive models when studying overweight and obesity are difficult to summarise, owing to their different socio-demographic and national contexts, as well as target groups (children, adolescents, or adults). Making methodological generalisations is also challenging. For example, the training set sizes in the studies have varied considerably – from more than 10,000 cases (eg. Kim et al. 2019; Pang et al. 2019) to a few hundred or even less (eg. Adnan et al. 2012; Rios-Julian et al. 2017). The same applies to the amount of input features in the models, generally ranging from ten to twenty, although models with roughly one hundred (eg. Pang et al. 2019) or hundreds of predictors (eg. Park et al. 2019) are also not uncommon.

Comprehensive overviews of the use of machine learning models in the study of overweight and obesity can be found, for example, in Chatterjee et al. (2020), Colmenarejo (2020), and Ferdowsy et al. (2021). Due to a lack of space, only a few studies will be introduced here and some of these will be returned to in the discussion section of the thesis.

- Cañas and Martinez (2020) employed supervised and unsupervised techniques of data mining, namely simple k-means, decision tree, and support vector machine to detect obesity levels among 178 students aged 18–25 in Colombia, Mexico, and Peru. The models had 18 input variables including various socio-demographic indicators, aspects of physical activity, dietary and health behaviour, alimentary disorders, and genetical factors. Of the two supervised algorithms, the decision tree method significantly outperformed the support vector machine model.

- Cheng et al. (2021) assessed the performance and predictive power of a set of common machine learning models, using data from 7,162 respondents to the National Health and Nutrition Examination Survey (2003–6) in the United States. Eleven classificatory algorithms were implemented, evaluated, and compared, including logistic regression, naïve Bayes, radial basis function, k-nearest neighbours, classification via regression, random subspace, decision table, multi-objective evolutionary fuzzy classifier, random tree, J48 (a type of a decision tree), and multilayer perceptron. When using altogether seven input variables (physical activity and six basic variables of demographic status), the random subspace classifier algorithm achieved the highest performance according to various conventional metrics. Logistic regression was middle-ranking in terms of overall accuracy, sensitivity, specificity, and the area under receiver operating characteristic curve.

- Curbelo et al. (2017) applied machine learning models when predicting obesity status based on genetic profile data publicly available in the National Human Genome Research Institute Catalog (United States). Seven machine learning models trained for the prediction of obesity were used, including gradient boosting, generalised linear model, classification trees, k-nearest neighbours, support vector machine, random forest, and multilayer perceptron. The models were first compared in terms of their ability to classify a subject into one of the BMI-related classes defined in the study based on the initial 6,622 variables describing genetic variants, age and gender. Dimensionality reduction resulted in an eventual set of 13 input variables with which the models performed far better than with all initial input features. Of the seven models evaluated, support vector machine emerged as the best-performing algorithm in this study.

- Delnevo et al. (2021) investigated the relationship between affect-related psychological variables and the BMI, using machine learning algorithms to forecast both BMI values and BMI status (normal, overweight, and obese) among 221 subjects. The algorithms employed included k-nearest neighbours, classification and regression tree, support vector machine, multilayer perceptron, adaptive boosting with decision tree, gradient boosting, random forest, and extra tree. All algorithms were used for both classification and regression problems. Additionally, for the regression analysis the authors also employed lasso and elastic net regression methods. The best performances were achieved by the extra tree classifier, although its performance was only slightly better than that of gradient boosting, random forest, and multi-layer perceptron. The study also confirmed that psychological variables, especially those of negative type (eg. depression) can be effectively used when predicting both the BMI values and the BMI status.

- Dugan et al. (2015) strove to predict childhood obesity after the age of two, using data collected prior to the second birthday among children belonging mainly to racial minority groups and low-income cohorts in the United States. Six different machine learning models were trained on a dataset of 7,519 cases and 167 variables: random tree, random forest, naïve Bayes, Bayes network, and two types of algorithms based on building a decision tree (J48 and ID3). Of these six models, ID3 turned out to have the best overall performance in terms of accuracy and sensitivity.

- Ferdowsy et al. (2021) assessed the performance of nine machine learning algorithms when predicting obesity in a sample of 1,100 Bangladeshis, using 28 input features in

their models that included k-nearest neighbours, random forest, logistic regression, multilayer perceptron, support vector machine, naïve Bayes, adaptive boosting, decision tree, and gradient boosting classifier. Based on various evaluation metrics, the overall performance was the best in case of logistic regression.

- Kim et al. (2019) analysed data from the 2017 Korean Youth Health Behaviour Survey, training ten machine learnings models (general Bayesian network, general Bayesian network with What-If analysis, logistic regression, decision tree, support vector machine, artificial neural network, naïve Bayes, bootstrap aggregating, random subspace, random forest) on the data from 11,206 respondents and while using 19 input variables. From these models, the general Bayesian network with What-If analysis had the best performance when predicting BMI categories.

- Pang et al. (2021) compared seven machine learning models: decision tree, Gaussian naïve Bayes, Bernoulli naïve Bayes, logistic regression, artificial neural network, support vector machine with radial basis function kernel, and XGBoost. These were trained to predict childhood obesity among children aged 2–7 (27,203 subjects), using healthcare records data collected before their second birthday by the Children's Hospital of Philadelphia. Four demographic and 54 clinical variables were included in the training of the models. XGBoost outperformed all other algorithms in terms of the area under receiver operating characteristic curve and achieved statistically significant better performance than other models also in terms of other standard classifier metrics.

- And finally, Thamrin et al. (2021) assessed the performance of logistic regression, classification and regression trees, and naïve Bayes when predicting obesity, trained on the data (618,898 cases) from the Indonesian Basic Health Research Survey. Logistic regression method had the best performance.

The core methodological aspects of these nine studies are summarised in Table 2 on the next page, also indicating the best performing machine learning algorithm according to the respective study. As can be seen from the table, the sample sizes in these studies, the amount of input variables included in the training of the models, and conclusions about model performance have varied considerably when predicting obesity (and in some cases also other weight categories or the actual value of the BMI).

Table 2

Examples of Studies on Overweight and Obesity Using Predictive Modelling

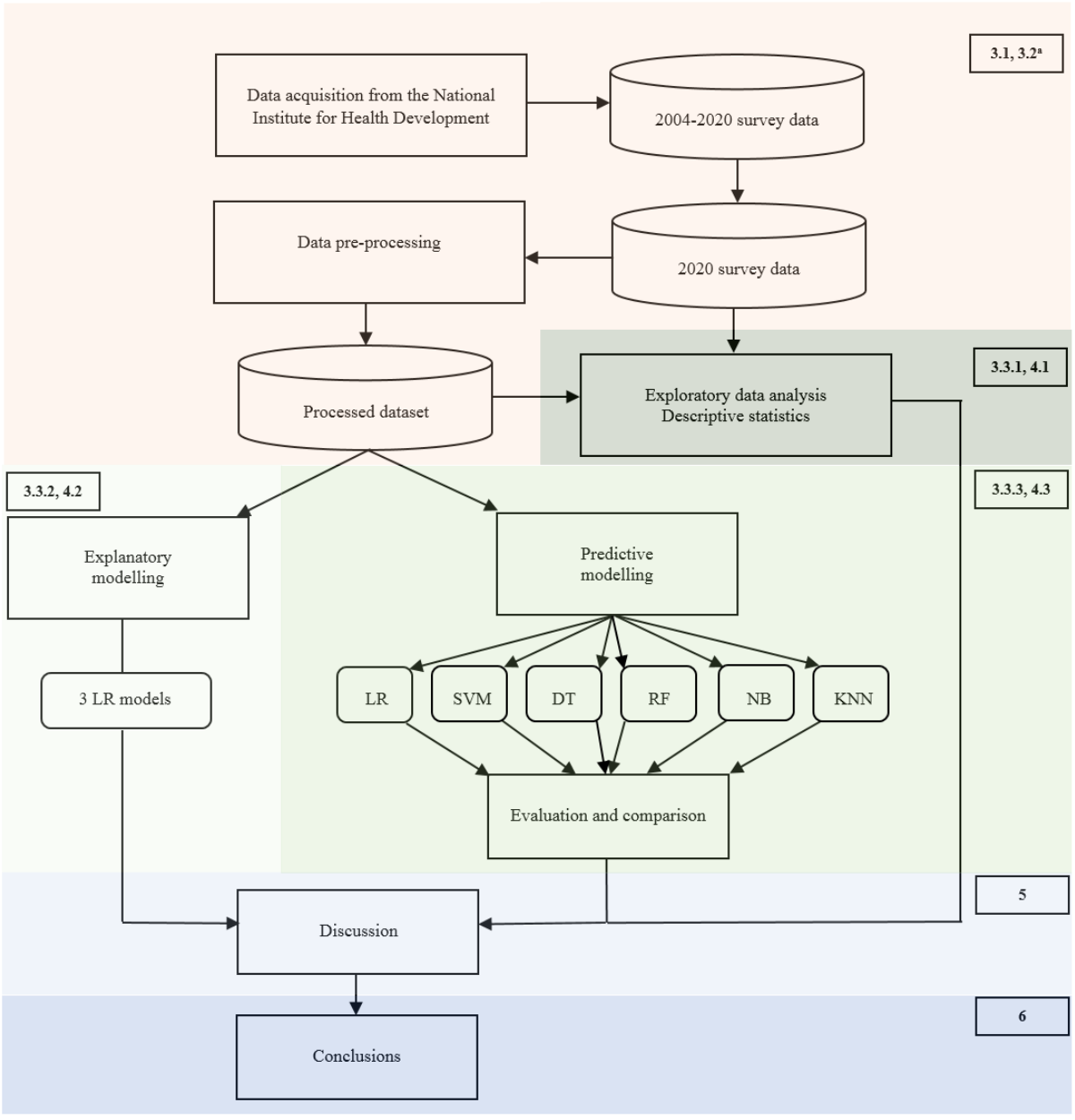
Study (year)	Country	Study group	Sample size	Output	n of input variables	Compared models[a]
Cañas and Martinez (2020)	Colombia, Mexico, Peru	18-25	178	Obesity	18	**DT**, SVM
Cheng et al. (2021)	US	All	7,162	Overweight, obesity	7	CVR, Dtable, J48, LR, KNN, MEFC, MLP, NB, RBF, RT, **RSS**
Curbelo et al. (2017)	ND	All	164	Obesity	13	CART, GB, GLN, KNN, MLP, RF, **SVM**
Delnevo et al. (2021)	Italy	Adults	221	BMI, Normal weight, overweight, obesity	10	AB, GB, RF, CART, ENT, **ET**, KNN, LASSO, MLP, SVM
Dugan et al. (2015)	US	< 2	7,519	Obesity	167	BN, **ID3**, J48, NB, RF, RT
Ferdowsy et al. (2021)	Bangladesh	ND	1,100	Obesity	28	ADA, GB, DT, KNN, **LR**, MLP, GNB, RF, SVM
Kim et al. (2019)	South Korea	12-18	11,206	Obesity	19	ANN, BA, DT, GBN, **GBN-MB**, LR, NB, RF, RSS, SVM
Pang et al. (2021)	US	2-7	27,203	Obesity	58	ANN, BNB, DT, GNB, LR, SVM, **XGB**
Thamrin et al. (2021)	Indonesia	18-	618,898	Obesity	21	CART, **LR**, NB

Note. Abbreviations: AB: Ada boosting; ANN: artificial neural networks; BA: bagging / bootstrap aggregating; BNB: Bernoulli naive Bayes; CART: classification and regression tree; CVR: classification via regression; DT: decision tree; Dtable: decision table; ENT: elastic net regression; ET: extra tree; GB: gradient boosting; GBN: general Bayesian network; GBN-MB: general Bayesian network with What-If analysis; GLN: generalised linear model; GNB: Gaussian naïve Bayes; ID3: iterative dichotomiser 3 (a form of a decision tree); J48: tree-based J48 model: k-NN: local k-nearest neighbours; LASSO: a method of linear regression; LR: logistic regression; MEFC: multi-objective evolutionary fuzzy classifier; MLP: multilayer perceptron; NB: naïve Bayes (distribution not specified); ND: not defined; RF: Random Forest; RSS: random subspace; RT: random tree; SVM: support vector machine: XGB: XGBoost

[a] The algorithm with the best performance is in bold.

3 Data and methods

This chapter gives an overview of the dataset used in this study and its origin, the steps of data pre-processing and feature selection principles, the nature of the performed statistical analysis, the chosen machine learning algorithms for predicting the BMI class and the rationale behind their choice, the principles of model training and testing, as well as the metrics used for model evaluation. Figure 2 below sums up in broad terms all the stages of data acquisition, pre-processing, and analysis process.



Note. Abbreviations: LR: logistic regression; SVM: support vector machine; DT: decision tree; RF: random forest; NB: naïve Bayes; KNN: k-nearest neighbours.

[a] The numbers in the boxes indicate the chapters or subchapters of the thesis where the respective stage is discussed.

Figure 2
Summary of the Research Process

Three types of software were used for data pre-processing and analysis: MS Excel (Office 365), IBM SPSS Statistics (Version 27.0.1), and Jupyter Notebook application in Anaconda distribution of Python 3.8.8. MS Excel was used for the initial major tasks of data pre-processing (feature elimination, generation of new variables such as BMI and BMI categories, and the rescaling of variables if deemed necessary). SPSS was used for binary logistic regression analysis and Jupyter Notebook was employed as the environment for evaluating and comparing the performance of different machine learning algorithms, as well as for data visualisation. Further refinements of the dataset were made also during data analysis stages, when necessary – with SPSS (transforming categorical variables into dummy variables for regression analysis) and with Jupyter Notebook (correlation analysis, dropping the rows with missing values, feature scaling and normalisation, and the transformation of categorical variables into dummy variables for machine learning models). Data pre-processing is described more in detail in section 3.2 below.

3.1 Description of the original dataset

The original dataset used for this study comprised the results for selected thematic blocks of the 2020 Health Behaviour Among Estonian Adult Population survey (from hereon HBEAPS). HBEAPS is a nationwide population-based cross-sectional survey, conducted biannually since 1990 by the Estonian National Institute for Health Development. It builds on a stratified random sample of 5,000 respondents (occasionally combined with additional regional samples) that is drawn from the Estonian resident population aged 16–64. The survey was part of the joint FinBalt Health Monitor project involving Finland and all three Baltic states until 2010, after which these countries have carried out the surveys independently (Reile and Veideman 2021, 1). The HBEAPS methodology has remained roughly the same over the years to ensure the comparability of the results, although the wording of various questions and their scales have changed. All biannual surveys have been approved by the Tallinn Medical Research Ethics Committee and carried out following the principles of best research practices and in accordance with the legislation that regulates the use of personal data (Reile and Veideman 2021, 2).

The main module of the HBEAPS covers various socio-demographic indicators and a wide range of topics related to the respondents' mental and physical health and behaviour, the use of healthcare services, smoking status and alcohol intake, nutrition, physical activity, and risk behaviour. The main module has occasionally been complemented with year-specific or rotating thematic blocks. All data collected by these surveys are based on self-reported answers to a questionnaire available in Estonian, Russian, and English, that is sent to the respondents included in the sample by mail and since 2016 the questionnaire can also be filled in online. Survey response rates have varied, ranging from as high as 77 percent in 1996 (Reile et al. 2020, 860) to 47.7 percent in 2020 (Reile and Veideman 2021, 7).

HBEAPS data are available for researchers and students for scholarly purposes and can be accessed through a process of formal application. The author of this thesis requested and received access to data on selected thematic blocks for the surveys of 2004-2020 (120-140 features, depending on the survey year; 25,113 respondents in total for all nine surveys). Because of the limited scope of this thesis, only the most recent HBEAPS data (n=1,737 after data pre-processing) are analysed here.[4] The 2020 HBEAPS questionnaire comprised 89

[4] Admittedly, a bigger dataset than this would have been beneficial, especially for predictive modelling. In some other studies that have analysed HBEAPS data, the results of different

questions, many of these having numerous sub-questions. The data were collected from an age- and gender-stratified random sample of 5,000 Estonian residents aged 16–64 drawn from those registered in the Estonian population register as of 1 January, 2020.[5] Similarly to previous surveys, due to the differences in response rates, female, more aged, and Estonian-speaking respondents are over-represented in the survey results compared to men, younger age groups and non-Estonians, taking into account their actual proportion in the Estonian adult population (Reile et al. 2019, 378). For this reason, the National Institute for Health Development has recommended that weight coefficients are used for gender and age groups when making inferences about the entire Estonian population based on the survey results (Reile and Veideman 2021, 11). In this thesis, weighted data were used in data exploration and logistic regression analysis stages, using appropriate sampling weights indicated in the National Institute for Health Development guidelines (Reile and Veideman 2021, 11). When comparing machine learning algorithms, unweighted data were used, since the aim of that endeavour was not to extrapolate the results to all Estonian population but simply to compare the performance of different models when predicting the BMI category.

3.2 Data pre-processing

The pre-processing of the data was carried out mainly in MS Excel (Office 365), and to a lesser extent in SPSS and Jupyter Notebook. It entailed relevant feature selection, generation of additional variables from the existing ones, the rescaling of the values for some variables, dropping cases with missing data, correlation analysis, and the creation of dummy variables. The initial dataset received from the Estonian National Institute for Health Development included 137 columns (i.e. variables) and 2,324 rows (i.e. cases). The first step in the data pre-processing stage was the generation of a new variable *BMI* calculated from the self-reported height and weight data by using the formula "BMI = weight (kilograms)/height^2 (meters squared)." The rows in the original dataset with unreported height, weight, or both were eliminated, reducing the size of the dataset at this point to 2,170 cases. For convenience and easy access in the data analysis stages, three additional variables were generated based on the just-calculated BMI values. Among these, the new variable *BMI_four_groups* grouped the respondents into four BMI categories based on the single-cut definition of these by the World Health Organization, as outlined in section 2.1 above.[6] Additionally, two dichotomous variables were generated. *BMI_two_groups_split25* divided the respondents into underweight/normal weight (BMI < 25.0 kg/m^2) and overweight/obese categories (BMI ≥ 25.0 kg/m^2). Variable *BMI_two_groups_split30*, in turn, divided the respondents into non-

---

survey years have been aggregated (eg. Reile and Leinsalu 2019). Such aggregation, however, was considered methodologically inappropriate for the purposes of this study. Aggregation of different biannual datasets based on random samples of 5,000 Estonian adult residents means that there exists a probably, however small, that the same person is included in the dataset more than once. Also, the socio-economic contexts in which the data for different surveys were collected, differ. This would be difficult to account for in the analysis and can thus lead to biased results.

[5] It is important to note that the data collection itself was carried out by the National Institute for Health Development from March to June of 2020, i.e. during the beginning of the COVID-19 pandemic which might have had at least some impact on the response rate as well as the nature of the responses.

[6] This variable was in the end excluded from the analysis, since the cases belonging to the underweight category (BMI < 18.0 kg/m^2) were very few (36) in the studied dataset and consequently binary instead of multivariate logistic regression models were employed.

obese (BMI < 30.0 kg/m^2) and obese categories (BMI ≥ 30.0 kg/m^2). These two dichotomous features were eventually used as dependent variables in this study.

The second step of data pre-processing included relevant feature selection from the existing 137 features to be used as independent variables in statistical analysis and as input variables in machine learning algorithms. In this study, for the sake of clarity, consistency, and comparability, the same set of features was used for both explanatory and predictive purposes. Feature selection and extraction to reduce the dimensionality of the data is a complex issue. Machine learning, owing to its focus on prediction rather than explanation, is generally regarded to be able to handle higher dimensionality than statistical modelling. As Shmueli (2010, 297) also points out, the criteria for choosing variables may differ considerably in explanatory and predictive contexts. In case of the former, where variables are seen as operationalised constructs, the choice of variables is generally based on the theoretical assumptions about their role in the putative causal explanation, while in case of the latter there exists no strong need for determining the exact role of each variable in terms of the underlying causal structure (*ibid*.). That being said, removing irrelevant and redundant data can increase learning accuracy of the predictive models and improve the comprehensibility of results, and is thus also important in machine learning (Khalid 2014) – at the very least for the purposes of avoiding model overfitting.

The features from the original dataset were eliminated iteratively, based on four principles, and roughly in the following temporal sequence:

- *Domain-specific assessment of feature relevance*. All features deemed as irrelevant for predicting the respondents' BMI category were omitted. An example: the original dataset included a question about recent dentist appointments and it was considered highly unlikely that this variable is associated with the respondent's BMI category.

- *Temporal reference to at least 12 months*. The questions in the HBEAPS questionnaire refer to different time periods (from one week to 12 months, depending on the question). From among potentially relevant questions that explicitly stipulated the time frame they referred to, only those questions were kept where the respondents were asked to self-report their behaviour during the previous 12 months. Questions referring to a shorter time period were not considered reliable indicators of a more long-term behaviour that might affect the BMI. An example: for this very reason, the independent variables in this study do not cover otherwise undisputedly relevant dietary behaviour because in the HBEAPS questionnaire, respondents are requested to self-report their dietary practices only for the "past 7 days."

- *Preference for objective over subjective responses.* Although all data collected by the HBEAPS questionnaire is self-reported, not objectively measured, the questions vary with regard to how subjective or objective the response to it potentially is. The questions leading to relatively more objective and concrete answers were prioritised. An example: the question on self-rated health status ("How would you assess your current state of health?") was omitted from the analysis, while a more factual question on suffering or not suffering from a chronic disease ("Do you have any longstanding (chronic) illnesses or health problems?") was retained. Another example: the question "How would you assess your current physical fitness/abilities?" was not included as a variable in the analysis, while the question "How often in your leisure time do you exercise for at least half an hour so that you will breathe a bit heavier and sweat a little?" was.

- *Avoidance of multi-collinearity*. After correlation analysis of various pairs or groups of features, only one of the correlated features was retained to reduce the potential multi-collinearity between eventually selected independent variables. An example: both employment status and marital status were strongly correlated with age, which is why both were excluded from the analysis. Another example: for the opposite reason, the selected features include both the status of suffering from a chronic disease and exercising frequency, although one of the categories of the latter in the questionnaire is "Cannot exercise due to injury or illness." No multi-collinearity between these two variables could be detected, probably because "chronic disease" refers to a wide variety of conditions, some more debilitating than others.

Based on these four principles, altogether ten features were finally selected as independent variables in this study. These comprised five socio-demographic (gender, age, ethnicity, education, and income) and five behavioural variables (having a chronic disease,[7] smoking status, alcohol intake,[8] exercising frequency, and physical effort required in one's everyday work).

The third stage of data pre-processing entailed the simplification of scales in case of some of the ten chosen variables by rescaling and relabelling them.[9] For example, in case of ethnicity, the HBEAPS questionnaire distinguishes between three classes ("Estonian," "non-Estonian," and "other"), but for the purposes of this study the variable was dichotomised by aggregating the classes "non-Estonian" and "other." The scales for education and income level,[10] alcohol intake, and exercising frequency were also simplified by aggregating several categories of the variable. Some scales (eg. for exercising frequency and smoking status) were reversed or the order of the categories was changed, so that in case of all ordinal variables, the numerical coding of the answers refers to an ascending order of the categories (i.e. from lower to higher

[7] The inclusion of "having a chronic disease" as a behavioural variable is, of course, debatable due to the underlying genetic causal factors in case of many chronic illnesses. However, the risks of developing a chronic disease or disorder can also be attributed to behavioural factors.

[8] The survey includes three potentially relevant questions on alcohol consumption: "In the past 12 months, how often did you consume alcoholic beverages?", "In the past 12 months, how often did you consume the following alcoholic beverages?" (reported separately for hard liquor, wine, beer, and other light alcoholic drinks), and "How often do you, at once, drink a) three bottles of (3x0.5 l, medium strong) beer or b) six glasses (6x120 ml) of wine or c) six shots (6x40 ml) of hard liquor?" The latter was chosen for this study and its preference owed to the question's undefined and hence more general temporal dimension. Other authors building on the HBEAPS data have used different strategies in this case. Reile and Leinsalu (2020, 1167), for example, created a binary variable on alcohol consumption aggregated from two different alcohol intake measures.

[9] Another reason for the rescaling was scale homogenisation across different HBEAPS survey years. Although this thesis focuses exclusively on the 2020 data, in further research the author intends to scrutinise more long-term trends of obesity as well as physical activity in Estonia, for which the homogenisation of scales is necessary.

[10] The variable on income (self-reported by respondents as their "family's average monthly income (income from all sources after tax) per family member") was the most difficult one to rescale and homogenise, since very different scales and income brackets with different widths have been used in different HBEAPS questionnaires. The eventual four categories of this variable correspond to the four quartiles of the scale used in the respective survey.

frequency, from less to more, etc). The initial variable *age* was replaced with the variable *age_group* for further simplicity and clarity when interpreting and making generalisations about the results of data analysis. The Estonian National Institute for Health Development in its reports on the HBEAPS data distinguishes between five age groups (16–24, 25–34, 35–44, 45–54, and 55–64). After some consideration, however, the first category (16–24) was left out from the current study to avoid the bias that this age group would have introduced into the analysis. More specifically, education and income levels – two among the ten features included in the analysis – would have been heavily biased for the respondents of this age group, most of whom are likely to be still in high school or university and thus have, almost by definition, lower level of education than most respondents belonging to other age groups, as well as little or no income.

Finally, all rows with any missing values were eliminated from the dataset reducing the size of it to altogether 1,737 cases. Removing rows with missing values is not always the best default action when cleaning data – it can introduce a bias to the analysis and there exist various data imputation measures as alternatives to the deletion of rows (Shmueli 2010, 296). However, since the values in this concrete dataset were missing randomly across all selected features, removing rows with any data missing was considered an appropriate course of action.

The descriptive characteristics of the final pre-processed dataset are summarised in Table 3 below.

Table 3

Descriptive Characteristics of the Dataset

Variables	Categories	Total (n=1,737)	
		n	%
Sociodemographic			
Gender	Male	692	39.8
	Female	1045	60.2
Age group	25-34	400	23.0
	35-44	432	24.9
	45-54	466	26.8
	55-64	439	25.3
Ethnicity	Estonian	1266	72.9
	Non-Estonian	471	27.1
Education	Primary or basic (1–9 years)	181	10.4
	Secondary (10–12 years)	292	16.8
	Secondary-vocational	497	28.6
	Higher	767	44.2
Income per household member	1st quartile	420	24.2
	2nd quartile	538	31.0
	3rd quartile	387	22.3
	4th quartile	392	22.6
Behavioural			
Chronic disease	Yes	794	45.7
	No	943	54.3
Smoking status	No	784	45.1
	Yes, used to smoke before	538	31.0
	Yes, currently occasionally	112	6.4
	Yes, currently every day	303	17.4

Alcohol intake	Never	928	53.4
	Less than once a month	427	24.6
	Once or a few times a week	164	9.4
	Once a week	117	6.7
	(Almost) every day	101	5.8
Exercising frequency	Cannot exercise due to injury or illness	158	9.1
	Once a month or less often	441	25.4
	2–3 times a month	135	7.8
	Once a week	253	14.6
	2–3 times a week	449	25.8
	4–7 times a week	301	17.3
Physical effort required in work	Very little	763	43.9
	Some	417	24.0
	Average	426	24.5
	A lot	131	7.5
BMI categories			
Four BMI groups	BMI < 18.5	36	2.1
	BMI = 18.5–24.9	742	42.7
	BMI = 25.0–29.9	575	33.1
	BMI ≥ 30.0	384	22.1
Two BMI groups (split at 25.0)	BMI < 25.0	778	44.8
	BMI ≥ 25.0	959	55.2
Two BMI groups split at 30.0)	BMI < 30.0	1353	77.9
	BMI ≥ 30.0	384	22.1

3.3 Methods of data analysis

This subsection gives a more detailed overview of the methods used at three distinct stages of data analysis (data exploration, explanatory modelling, and predictive modelling) in this study.

3.3.1 Exploratory data analysis

In the exploratory data analysis phase, descriptive statistics were used to discern basic information about the prevalence of overweight and obesity by features included in the dataset. This entailed calculating cross-sectionally the stratified means and standard deviations of the BMI and the proportions of overweight (BMI ≥ 25.0 kg/m^2) and obese (BMI ≥ 30.0 kg/m^2) population for each socio-demographic and behavioural feature. The statistical significance of the differences between the means of the BMI across the categories of each variable was tested with the independent samples t-test for binary variables such as gender, ethnicity, and chronic disease and with One-Way ANOVA for the analysis of variance in case of age group, education and income level, smoking status, alcohol intake, exercising frequency, and the physical effort required in work. When analysing variance, Tukey's b and Games-Howell Post-Hoc tests for pairwise multiple comparison were also applied to test for the difference between each pair of means. Since all socio-demographic and behavioural features in the analysed dataset are categorical, Chi-square test was used to examine the statistical significance of the associations between belonging to the overweight or obese category and the respective socio-demographic and behavioural features. In all statistical tests applied, the level of significance was set at $p < 0.05$.

Both SPSS version 27.0.1 and Jupyter Notebook application in Anaconda distribution of Python 3.8.8 were employed at this stage of data analysis. Descriptive statistics on the

stratified proportions of overweight and obese population by feature were selectively visualised by creating bar and density plots from the explored data in Jupyter Notebook.

3.3.2 Explanatory modelling

Explanatory modelling entailed studying the associations between various socio-demographic and behavioural variables on the one hand, and weight categories on the other by means of binary logistic regression. When performing the regression analysis, all categorical variables with more than two classes were first transformed into dummy variables, with the lowest level set as the baseline reference category. The strength of the associations between overweight (BMI $\geq$ 25.0 kg/m^2) and obesity (BMI $\geq$ 30.0 kg/m^2) class and the independent variables was estimated by odds ratios (OR) and reported with 95% confidence intervals (CI) for each class of the independent variable.

Three different models were run, separately for both overweight and for obesity as the dependent variable, to test also for the possible differences between these weight groups in terms of their association with the selected independent variables. The first model focused on each of the ten independent variables separately to assess their association with either overweight or obesity. The model was thus run, in fact, ten times for both dependent variables and the crude odds ratios with 95 percent confidence intervals for each independent variable are reported in the column "Model 1" in Tables 7 and 8 in section 4.2. The second model consisted in essence of a set of six submodels – the first adjusted simultaneously for all socio-demographic variables, and the other five for all socio-demographic variables and one behavioural variable at a time. The odds ratios with 95 percent confidence intervals for Model 2 are reported in the respective column in Tables 7 and 8 in section 4.2 – the figures for socio-demographic variable are the results of the first submodel, the figures for behavioural variables are the results of submodels adjusted for five socio-demographic variables and the respective behavioural variable. The third model was simultaneously adjusted for all ten independent variables in the study (i.e. both the five sociodemographic variables and the five behavioural variables) in order to assess the effect of each variable on overweight and obesity status while accounting for the effects of all other variables.

The default "enter" method of regression was used in all models, meaning that all adjusted for independent variables were included into the model in a single step, and the eventual model did not eliminate the statistically insignificant variables, unlike in case of "forward or backward stepwise" regression. This method was chosen in order to be able to report the results for all independent variables, also the ones that did not improve the model fit or where the associations were not statistically significant. Regression analyses were performed with IBM SPSS Statistics software (version 27.0.1).

3.3.3 Predictive modelling

Predictive modelling focused on evaluating and comparing the performance of six supervised machine learning algorithms that are considered to be particularly suitable for classification problems and that have been commonly used also in other studies when predicting overweight or obesity (see section 2.4).

Compared models

The list below is a brief description of these algorithms and the rationale behind their choice in the analysis.

- *Decision tree (DT)* is a predictive classification method with the goal of assigning samples to specific classes while determining the probability thresholds derived from input data (DeGregori et al. 2017, 672). As Colmenarejo (2020, 9) explains, the idea is "to generate rectangular partitions of the space of predictor variables, by successive splitting of the data by (usually binary) splits in one variable that optimises some loss function." Falling above or below the threshold moves the datapoint into the appropriate class and divides the data recursively into increasingly smaller subdivisions (*ibid*.). The resultant decision tree is composed of a root node, a set of internal nodes generated by the data division and terminal nodes that correspond to the results of the decisions (*ibid*.). Decision tree method can be implemented through different algorithms such as CART, ID3, J48, CHAID, and many others. The upside of the decision tree method is its simple and intuitive structure and thus high interpretability (Breiman 2001, 207; Colmenarejo 2020, 9), and its capacity to take into account nonlinear relationships among multiple covariates (Doupe et al. 2019, 811). The downside, however, is that decision trees are not necessarily "great predictors" (Breiman 2001, 207), and some types of decision trees suffer from selection bias (Kern et al. 2019, 3) or are prone to overfitting (Doupe et al. 2019, 811).

- *Random forest (RF)* is a common example of ensemble methods, where a model of higher quality is built by aggregating multiple models of lower quality (Colmenarejo 2020, 10). The prediction for new instances in ensemble methods is obtained either by averaging the prediction of all simpler models in case of regression problems or by majority voting in case of classification problems (*ibid*.; Breiman 2001, 207; Doupe et al. 2019, 811). Such predictions are generally more robust and are likely to have higher accuracy. In case of random forest, an ensemble of multiple, possibly hundreds or even thousands of decision trees are grown successively, often by using the so-called "bagging approach" (Colmenarejo 2020, 10). The sampled data are handed over to a CART-like algorithm in order to grow a decision tree on each bootstrap sample, respectively, and since these samples comprise different portions of the original data, the corresponding trees tend to differ across samples and therefore form an ensemble of distinct trees (Kert et al. 2019, 8). The advantage of random forest over single decision trees is that it is generally a better predictor, although its mechanism for producing a prediction can be difficult to grasp intuitively (Breiman 2001, 208). Like decision trees, random forest requires little data pre-processing, the features do not need to be normalised and feature selection procedures are not required since the algorithm does this on its own (Ahmadi et al. 2020, 5; Doupe et al. 2019, 811)

- *Support vector machine (SVM)* builds a so-called "hyperplane" from the predictor variables with maximal margin (Colmenarejo 2020, 9). "Hyperplane with a maximal margin" in this context refers to a hyperplane that has the largest distance to the training instances from an infinite number of possible hyperplanes (*ibid*.). The hyperplane itself is defined by the closest points to it called the support vectors from which the method takes its name (*ibid*.). If data are linearly separable, the support vector machine is based on the distance values between the hyperplane and the two

data classes (Ahmed and Loutfi 2013, 5). In case of data not separable in a linear manner, a distinct kernel function can be applied of which there are many types (*ibid*.). Support vector machine algorithms are considered to be particularly robust in case of binary classification, for which, in fact, they were originally designed although later developments of the method have allowed them to deal also with multiclass classification as well as regression (*ibid*.).

- *Naïve Bayes (NB)* classification algorithms belong to a bigger family of Bayesian classifiers that predict class membership based on probabilities, such as the probability that a given sample belongs to a particular class (Zhang 2009, 452). Naïve Bayes assumes that the effect that an attribute plays on a given class is independent of the values of other attributes (*ibid.*). The classification by the algorithm is achieved by applying the Bayes rule to calculate the probability of each attribute and by predicting the class based on the highest prior probability (Thamrin et al. 2021, 4). The advantage of the Bayesian family of algorithms is considered to be their high level of accuracy and speed when using large data sets (*ibid*.).

- *K-nearest neighbors (KNN, also k-NN)* classification algorithm assigns a class label to new instances based on the class assignment of their k nearest neighbours with predictor variables least distant from those of each new instance. This distance can be measured by various metrics, Euclidean distance being the most common one (Colmenarejo 2020, 8). Like in case of random forest algorithm, the class assignment decision is made by majority voting and the decision about continuous labels by the (weighted) average of the labels of the k-nearest neighbours (*ibid*.). The value of k can vary considerably depending on the dataset and can be determined by trial and error or cross-validation techniques.

- *Logistic regression*, also employed as a "traditional statistical method" at the explanatory modelling stage in this thesis, is one of the most common classification models to predict binary outcomes or multi-class outcomes in case of multivariate logistic regression (Thamrin et al. 2021, 3). The model takes a form of an intuitively easy to grasp linear equation to predict the log-odds of a binary variable (Colmenarejo 2020, 14).

Model training, testing and validation methods

Model training, testing, and validation methods followed the general conventions of machine learning in case of all six algorithms. The dataset was randomly divided into a training set composed of 80 percent (n =1,389) and a test set composed of 20 percent of the cases (n = 348). For comparative purposes, ten-fold cross-validation method was also tried out but in case of predicting overweight in particular, it did not lead to considerably different scores from those achieved by train/test split method.

The dataset used in this study is fairly balanced in terms of the two classes of the output variable *BMI_two_groups_split25* (44.8 percent of the cases belonging to the class "BMI < 25.0 kg/m^2" and 55.2 percent of the cases belonging to the class "BMI ≥ 25.0 kg/m^2"). The data are, however, unbalanced in case of the variable *BMI_two_groups_split30* – 77.9 of the cases in the pre-processed data belong to the class "BMI < 30.0 kg/m^2" and 22.1 percent to the class "BMI ≥ 30.0 kg/m^2". Such imbalance could lead to predictions that are more accurate on the majority class than on the minority class, thus resulting in a bias in favour of the former. To handle the class imbalance when predicting obesity, the resampling technique

SMOTE was used to over-sample the minority class by generating synthetic cases belonging to this class in the training set and thus securing that the number of cases with obese and non-obese status was the same and the training data hence balanced.

The value of the random state in all models was set to "1" to ensure the reproducibility of results. The primary aim of predictive modelling was to compare – in general terms – the performance of different algorithms, especially with logistic regression, also used in the explanatory modelling stage, as well as the potential differences between the performance of these algorithms when predicting overweight and obesity. The aim, however, was not to find the most fine-tuned version of each algorithm through further calibration and optimisation of their hyperparameters. For that reason, default hyperparameters were generally used in case of all algorithms, with the exception of support vector machine where the kernel function was set to "linear" and the probability function to "True" in order to enable the determination of feature importance and the plotting of the ROC curve. The optimal k value for the k-nearest neighbors algorithm was determined by testing and plotting the accuracy of k in the range of 1–40, resulting in k=31 for predicting overweight and k=3 for predicting obesity.[11]

The following Table 4 sums up the compared algorithms and their hyperparameters.

Table 4

The Hyperparameters of the Compared Machine Learning Algorithms

Model	Hyperparameters
LR	'C': 1.0, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'random_state': 1, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False
DT	'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'random_state': 1, 'splitter': 'best'
RF	'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'auto', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_jobs': None, 'oob_score': False, 'random_state': 1, 'verbose': 0, 'warm_start': False
SVM	'C': 1.0, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 'scale', 'kernel': 'linear', 'max_iter': -1, 'probability': True, 'random_state': 1, 'shrinking': True, 'tol': 0.001, 'verbose': False
NB	'priors': None, 'var_smoothing': 1e-09
KNN	'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': None, 'n_neighbors': 31[a] or 3[b], 'p': 2, 'weights': 'uniform'

Note. [a] prediction of overweight; [b] prediction of obesity

[11] The rationale behind the method of finding the optimal k value was to first determine and plot the accuracy of all k values up to roughly the square root of the training set size and then to choose the odd k value (to avoid the situation of two class labels achieving the same score) with the highest testing accuracy.

The same rationale – i.e. aiming at high-level comparison instead of algorithm optimisation – also defined how the predictor variables were dealt with. Optimisation of machine learning models often includes recursive elimination of features to determine the most optimal set of predictors. Such process simplifies the pool of estimators and penalises complexity in case of correlated covariates by choosing some and dropping others (Doupe et al. 2019, 809). In this study, however, feature selection was carried out already in the data pre-processing phase, the same set of features was purposefully used in both explanatory and predictive modelling stages, and further recursive elimination of features in case of predictive modelling was not applied in order to render the two modelling approaches as comparable as possible.

Ten input features in total might sound like a meagre set for machine learning purposes, considering that it is supposed to be able to handle multi-dimensional, complex, and "wide" data. However, most other studies that were introduced in section 2.4 similarly operated with a relatively limited set of 10–20 predictors. No prior feature selection can lead to model overfitting, since in high-dimensional data, not all features are equally relevant. In fact, some or many features might not be relevant for the training of the model at all. The latter, as Khalid et al. (2014) categorise them, are features that are either "irrelevant", "redundant," or outright "misleading." Curbelo et al. (2017, 4), for example, describe their attempt to predict obesity based on the initial 6,622 features – data on 6,620 single-nucleotide polymorphisms (SNPs), and age and gender. This analysis led to very poor, practically random results, and the authors cut down the amount of eventual input variables to thirteen features, based on both performing feature selection using random forest algorithm and their own domain-specific knowledge. As Curbelo et al. (*ibid*.) conclude, machine learning models may suffer considerably from the decrease in their performance when the number of features is excessively high.

Model evaluation and comparison

To evaluate and compare the performance of these six models when predicting both overweight and obesity, the following conventional evaluation metrics were used: recall (or sensitivity), precision, accuracy, and F1-score. The computation of all these measures is based on a confusion matrix of actual and predicted values (Buskirk et al. 2018). Additionally, the metric of area under receiver operating characteristic curve (AUC), also known as C-statistic (Doupe et al. 2019, 813), was used. The following is a brief explanation of these five metrics all of which are suitable for the evaluation of binary classification models. The abbreviations in the equations stand for the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).[12]

- *Recall*, also known as sensitivity or the true positive rate, is the proportion of true positives that are correctly identified by the classifier from all positive cases. The respective equation for calculating sensitivity:

$$Recall = \frac{TP}{TP + FN}$$

- *Precision* refers to the proportion of true positives among all instances predicted as positive by the classifier. The respective equation for calculating precision:

[12] These equations are used to calculate either fractions or percentages – in the latter case the calculations presented here need to include multiplication by 100.

$$Precision = \frac{TP}{TP + FP}$$

- *Accuracy* is essentially the proportion of all cases predicted correctly from the total number of predictions. The respective equation for calculating accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- *F1-score* (also known as *F-score*) is the measurement of the harmonic mean of recall and precision. It is considered to be more robust a metric than overall accuracy, since it is less affected by class size imbalances (Ahmadi et al. 2020, 5). The respective equation for calculating F1-score:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- The receiver operating characteristic (ROC) curve is a curve drawn in a two-dimensional plot where the x axis indicates the false positive rate and y axis the true positive rate. The *area under the curve* (AUC) value can be calculated from this curve. A higher value the AUC statistic indicates a better performance of the classification model. The values that are close to 0.5 indicate poor model performance, values that are closer to 1, on the contrary, refer to more accurate models (Buskirik et al. 2018, 6).

Apart from these metrics, the feature importances when predicting overweight and obesity status were determined by using either *feature_importances_* or *coef_* attribute in Python's sklearn, depending on the predictive model analysed. The exception was the k-nearest neighbors algorithm, in which case a direct calculation of feature importances or coefficients is not possible.

Predictive modelling was performed with Jupyter Notebook application in Anaconda distribution of Python 3.8.8. The specific Python libraries used in this process are listed in the following table:

Table 5

Python Libraries Used in Data Analysis

Python library	Purpose of use
pandas	Data importing, structuring, and analysis
numpy	Computations with multi-dimensional array objects
scipy	Scientific and technical computing
matplotlib	Data visualisation and graphical plotting
seaborn	Data visualisation and graphical plotting
scikit-learn /sklearn	Machine learning, cross-validation, and model evaluation
imblearn	Dealing with imbalanced classes

The Python scripts used for analysis are available in GitHub at this link: https://github.com/toomasgross/thesis

4 Results

This chapter offers an overview of the results of data analysis and is divided into three subsections that correspond to the stages of the analysis.

4.1 Descriptive statistics

The weighted cross-sectional differences of mean BMI (reported together with standard deviation), overweight and obesity status are displayed in Table 4. As can be seen from the results, men were more overweight and more obese than women (66.4 compared to 45.9 and 22.6 compared to 20.2 percent, respectively), although the difference was statistically significant only in case of overweight. There were statistically significant differences in the prevalence of overweight and obesity between all four age groups as pair-wise comparison with post-hoc tests also revealed. The higher the age group, the higher the prevalence of overweight and obesity. Non-Estonians were relatively more overweight than Estonians (61.0 compared to 51.4 percent). Estonians, in turn, were relatively more obese than non-Estonians (21.7 compared to 20.7 percent, respectively), although the difference was not statistically significant. Statistically significant differences also existed in the prevalence of overweight and obesity by level of education and income, although Tukeys'b and Games Howell tests revealed that only the mean BMI of respondents with higher education and those in the first income quartile was different from that in all other classes of the variable at 0.05 level of significance. The population with higher education was the least overweight and obese (46.6 and 15.3 percent, respectively), while the respondents with secondary-vocational education tended to be the most overweight (66.2 percent), and those having only primary or basic education (up to 9 years of schooling) tended to be the most obese (28.6 percent). In case of income, the differences between the proportions of overweight population in the four quartiles were statistically significant, unlike in case of obesity. Respondents in the first income quartile were relatively most obese (28.5 percent), while those in the fourth quartile were the least obese (15.9 percent).

In case of behavioural variables, the differences in the mean BMI between classes were statistically significant ($p < 0.05$) in all five cases. Respondents who had a chronic disease were relatively more overweight and obese than those who did not (64.4 compared to 49.6 and 31.2 compared to 13.5 percent, respectively). Differences by smoking status were also statistically significant – for both overweight and obesity. Interestingly, former smokers tended to be more overweight and obese (61.2 and 26.2 percent, respectively) than all other groups. Higher levels of alcohol consumption were mostly associated with higher proportions of overweight and obesity among the respondents. The most overweight (70.9 percent) and obese category (28.2 percent) in case of alcohol intake were the ones who consumed alcohol (almost) daily. In case of exercising frequency, the mean BMI of the category that did not exercise at all due to injury or illness was statistically significantly different from that of all other categories. These respondents were by far the most overweight (75.8 percent) and obese (39.0) from among the categories of this variable. And finally, the physical effort required in daily work appeared, perhaps counterintuitively, to be positively associated with overweight and obesity prevalence. The more physically demanding the work, the higher the prevalence of overweight and obesity, and these differences between categories were statistically significant.

Table 6

Mean BMI, Overweight and Obesity Prevalence by Classes of Different Variables

Variable	Class		BMI		Percentage	
			M	*SD*	Overweight	Obese
Gender	Male [a]		27.3	4.8	66.4	22.6
	Female [a]		25.8	5.9	45.9	20.2
		p	*<.001*		*<.001*	*.221*
Age group	25-34[a]		24.6	4.5	39.0	9.8
	35-44[a]		25.9	4.9	53.5	16.8
	45-54[a]		27.4	6.0	61.8	24.9
	55-64[a]		28.7	5.5	72.8	36.3
		p	*<.001*		*<.001*	*<.001*
Ethnicity	Estonian		26.5	5.6	54.4	21.7
	Non-Estonian		26.7	5.0	61.0	20.7
		p	*.161*		*.014*	*.667*
Education	Primary or basic (1–9 years)		27.7	5.6	62.8	28.6
	Secondary (10–12 years)		26.8	5.0	59.5	21.0
	Secondary-vocational		27.7	5.9	66.2	28.1
	Higher [a]		25.4	5.0	46.6	15.3
		p	*<.001*		*<.001*	*<.001*
Income per household member	1st quartile [a]		27.4	6.6	59.3	28.5
	2nd quartile		26.4	5.2	58.0	21.0
	3rd quartile		26.4	5.1	55.1	20.4
	4th quartile		26.0	4.7	52.0	15.9
		p	*<.001*		*.152*	*<.001*
Chronic disease	Yes [a]		27.9	6.3	64.4	31.2
	No [a]		25.5	4.4	49.6	13.5
		p	*<.001*		*<.001*	*<.001*
Smoking status	No		25.9	5.3	50.9	18.3
	Yes, used to smoke before		27.4	5.9	61.2	26.2
	Yes, currently occasionally		26.4	5.1	59.3	21.1
	Yes, currently every day		26.7	4.8	59.0	20.8
		p	*<.001*		*<.001*	*.008*
Alcohol intake	Never		26.3	5.6	52.1	20.8
	Less than once a month		26.3	5.5	54.0	18.1
	Once or a few times a week		27.3	5.1	63.9	23.3
	Once a week		27.5	5.0	66.2	27.6
	(Almost) every day		27.5	4.6	70.9	28.2
		p	*.004*		*<.001*	*.042*
Exercising frequency	Cannot exercise due to injury or illness [a]		29.7	7.6	75.8	39.0
	Once a month or less often		26.7	5.4	54.5	24.8
	2–3 times a month		26.2	4.2	58.5	17.8
	Once a week		26.4	5.1	57.1	22.4
	2–3 times a week		25.8	4.7	52.3	16.8
	4–7 times a week		26.1	5.3	52.6	15.3
		p	*<.001*		*<.001*	*<.001*
Physical effort required in work	Very little		26.3	5.7	52.0	19.4
	Some		26.3	5.2	56.1	18.9
	Average		26.9	5.3	61.0	23.7
	A lot		27.6	5.3	63.9	32.4
		p	*.023*		*.005*	*.002*

Note. Statistical significance of the differences between mean BMI in different classes was determined through the independent samples t-test (binary variables) and one-way ANOVA (non-binary variables). In case of ANOVA, post-hoc tests (Tukey's b and Games Howell) were used to pairwise compare the differences between classes. The statistical significance of the differences between the proportions of overweight and obese populations in each class was tested with Pearson's Chi-squared test.

[a] The mean BMI of the class is different from that in all other classes at the 0.05 level of significance.

The following two figures (3 and 4) help to better visualise the above-described differences between the categories of socio-demographic and behavioural variables, except for age that is presented separately on figures 5 and 6.

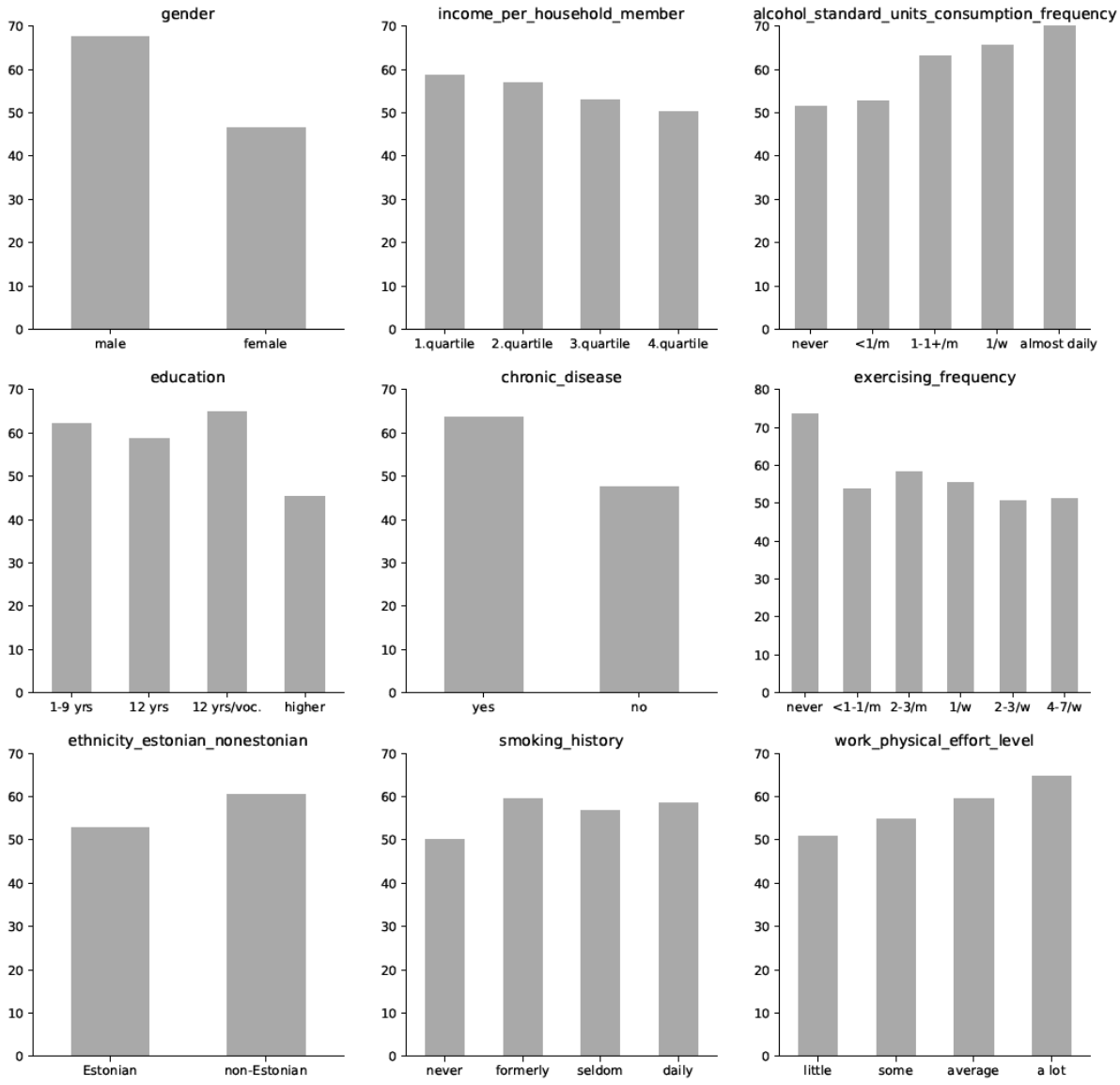


Figure 3
Percentage of Overweight Population by Variable Classes

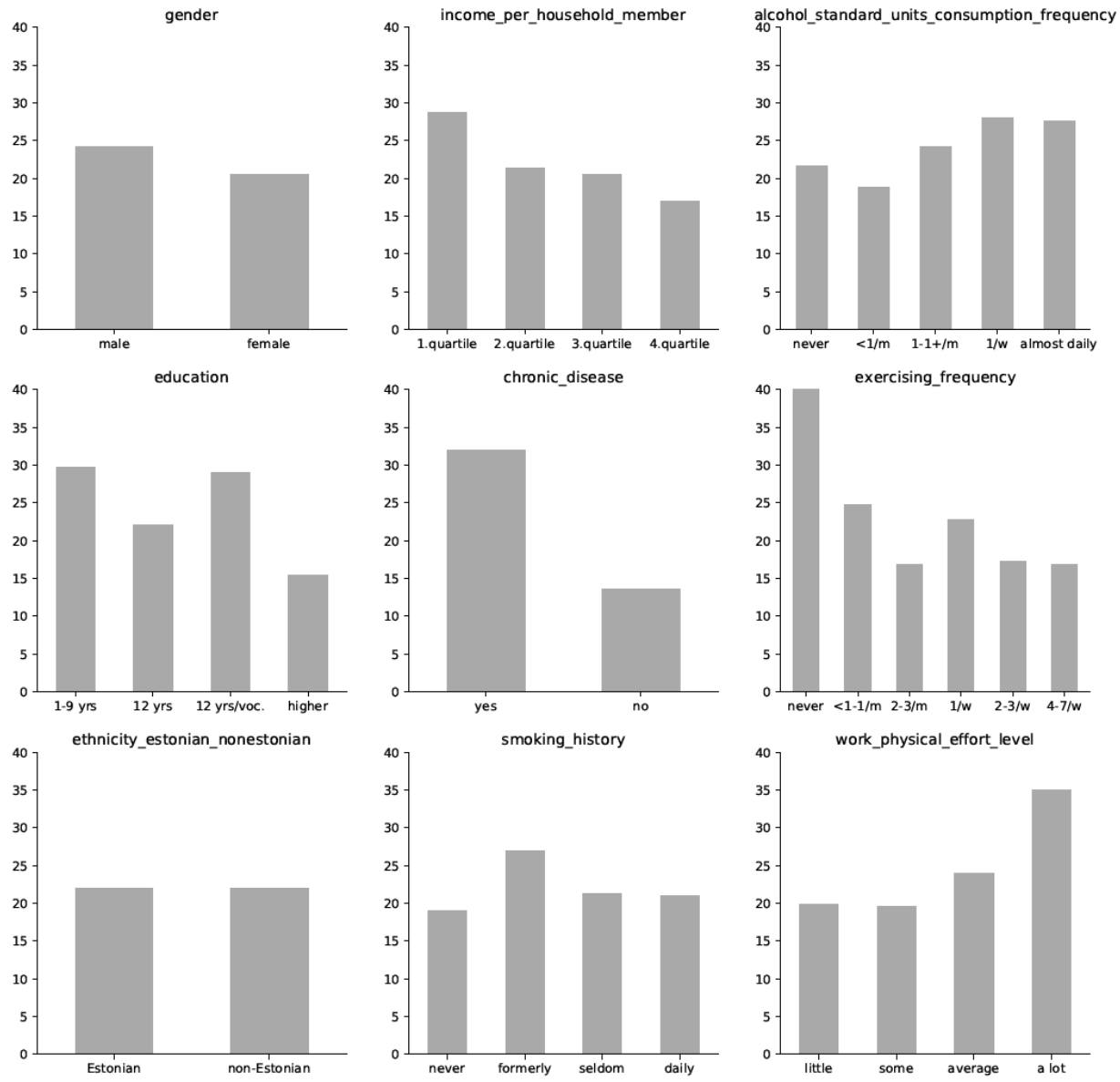
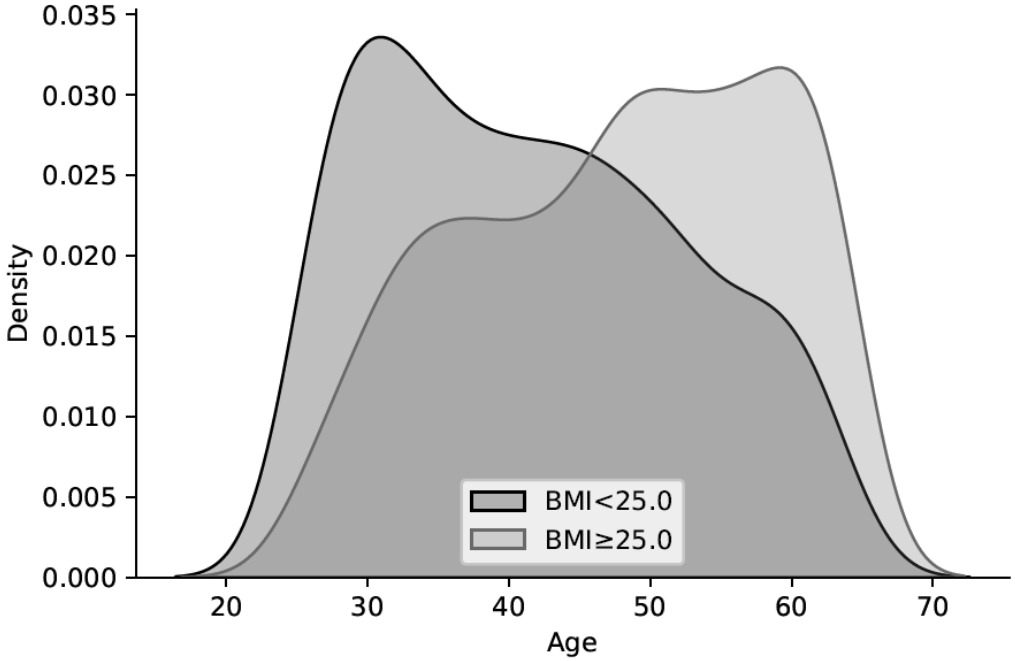


Figure 4
Percentage of Obese Population by Variable Classes

The following two density plots (figures 5 and 6), in turn, demonstrate the age distributions of respondents belonging to different weight categories (non-overweight and overweight, and non-obese and obese, respectively). As can be seen from these figures, the distributions of non-overweight and non-obese population are right-skewed (in other words, skewed positively), whereas the distributions of overweight and obese population are left-skewed (i.e. skewed negatively), and the skewedness is particularly marked in case of non-overweight (BMI < 25.0 kg/m^2) and obese categories (BMI ≥ 30.0 kg/m^2). This fact further underlines the result of exploratory data analysis reported above that age seems to be strongly associated with overweight and obesity status.



Note. The density plot approximates the underlying distribution, also outside the data range (ages 25-64), which is why the density values of the plot extend beyond the minimum and maximum age in the actual data.

Figure 5
Age Distribution of Non-overweight and Overweight Population

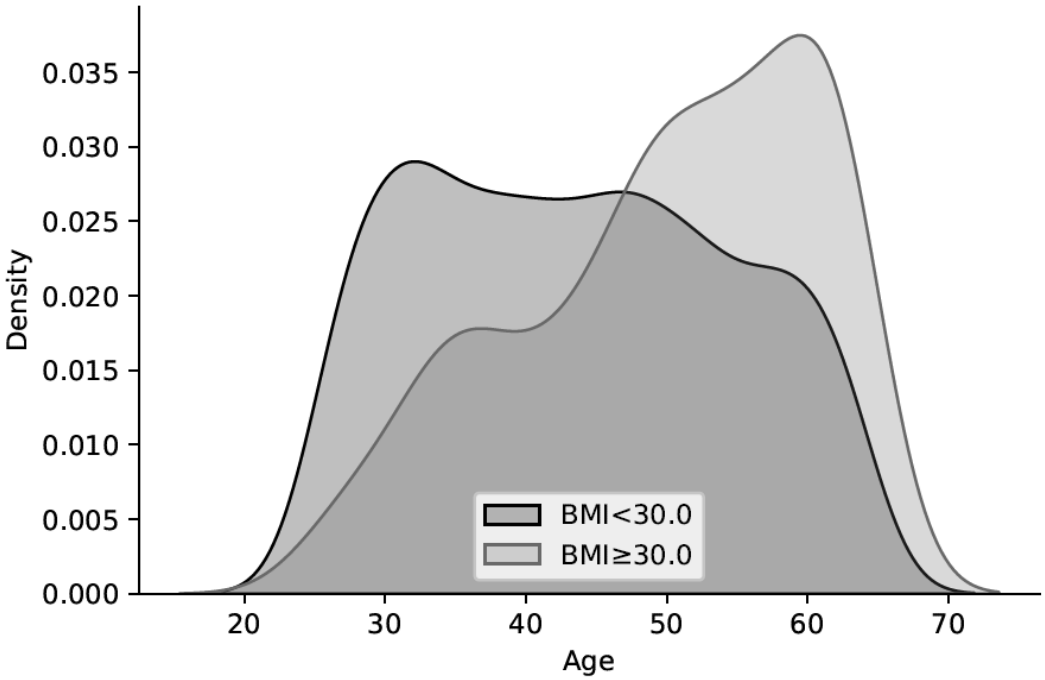


Figure 6
Age Distribution of Non-obese and Obese Population

4.2 Explanatory modelling

Tables 5 and 6 display the results of binary logistic regression analyses for the associations between overweight and obesity status and ten self-reported socio-demographic and behavioural variables. Three regression models were run for both overweight and obesity. Model 1 simply estimates the association between the weight category as the dependent variable and each independent variable separately, and the corresponding column in the table indicates the crude odds ratios for each independent variable. Model 2 was run to estimate the associations between different sets of variables and weight categories. The figures for sociodemographic variables refer to the odds ratios in a model adjusted only for these five sociodemographic variables. The figures for behavioural variables under Model 2 refer to odds ratios in models adjusted for all five socio-demographic variables and the respective behavioural variable. Model 3 was mutually adjusted for all ten independent variables included in this study (i.e. the five sociodemographic variables and the five behavioural variables).

As can be seen from the results, in case of Model 1, statistically significant associations could be detected between overweight status and at least one category of all independent variables. However, when mutually adjusted for five socio-demographic variables (Model 2) or for all independent variables included in this study (Model 3), various among these associations became statistically insignificant. The estimates in Model 3 – the most "holistic" of the models – deserve a slightly closer scrutiny here.

Associations between being in the overweight category and the socio-demographic or behavioural feature were most consistent for the variables of gender, age group, ethnicity, and chronic disease. Incidentally, this was so across all three regression models. According to Model 3, women had considerably lower odds of being overweight than men (OR 0.42, 95% CI 0.33–0.54), and those having no chronic disease had lower odds of being overweight than those who suffered from such a disease (OR 0.71, 95% CI 0.57–0.89). Conversely, non-Estonians had higher odds of being overweight than Estonians (OR 1.24, 95% CI 0.97–1.58). The statistically significant associations were the most consistent, however, for age group. The respondents in the highest age group (aged 55–64), for example, were 4.41 times (95% CI 3.17–6.14) more likely to be overweight than those in the lowest age group (aged 25–34), the baseline reference category. Statistically significant inverse associations were also observed for some categories of education and exercising frequency. Respondents with higher education had far lower odds of being overweight than those with primary or basic education, the baseline reference category of this variable (OR 0.56, 95% CI 0.38–0.82). In case of exercising frequency, those respondents who exercised 4–7 times a week (OR 0.6, 95% CI 0.37–0.98) and also those who exercised once a month or less often (OR 0.59, 95% CI 0.37–0.93) had lower odds of being overweight compared to the baseline reference category, i.e. those who did not exercise at all. Somewhat surprisingly, no statistically significant associations were detected for any other categories of exercising frequency – although the odds of being overweight were lower for all exercising categories compared to the baseline category, the estimates of this were not significant at $p < 0.05$ level. In case of alcohol intake, statistically significant association with overweight status was detected only for the group who consumed alcohol (almost) every day. Their odds of being in the overweight category were 1.63 (95% CI 1.00–2.65) compared to those who never consumed alcohol.

No statistically significant associations were present for any categories of income level, smoking status, and physical effort required by daily work.

Table 7

Associations between Socio-demographic and Behavioural Variables and Overweight Status Reported as Odds Ratios with 95% Confidence Intervals

Variable	Category	Model 1		Model 2		Model 3	
		OR	95% CI	OR	95% CI	OR	95% CI
Sociodemographic							
Gender	Male	1		1		1	
	Female	**0.43**	**0.35–0.52**	**0.39**	**0.32–0.49**	**0.42**	**0.33–0.54**
Age group	25-34	1		1		1	
	35-44	**1.81**	**1.39–2.35**	**1.91**	**1.45–2.53**	**1.93**	**1.45–2.56**
	45-54	**2.53**	**1.93–3.33**	**2.73**	**2.03–3.65**	**2.68**	**1.98–3.62**
	55-64	**4.21**	**3.15–5.63**	**4.7**	**3.44–6.41**	**4.41**	**3.17–6.14**
Ethnicity	Estonian	1		1		1	
	Non-Estonian	**1.31**	**1.05–1.63**	1.21	0.95–1.53	**1.24**	**0.97–1.58**
Education	Primary or basic (1–9 years)	1		1		1	
	Secondary (10–12 years)	0.87	0.60–1.26	0.84	0.56–1.24	0.83	0.56–1.25
	Secondary-vocational	1.16	0.82–1.63	1.01	0.70–1.46	1.03	0.70–1.51
	Higher	**0.52**	**0.37–0.71**	**0.58**	**0.41–0.82**	**0.56**	**0.38–0.82**
Income	1st quartile	1		1		1	
	2nd quartile	0.95	0.73–1.24	1.11	0.84–1.48	1.13	0.84–1.51
	3rd quartile	0.84	0.63–1.11	1.08	0.79–1.47	1.15	0.83–1.58
	4th quartile	**0.75**	**0.56–0.98**	1.07	0.78–1.47	1.11	0.80–1.54
Behavioural							
Chronic disease	Yes	1		1		1	
	No	**0.54**	**0.45–0.66**	**0.69**	**0.55–0.85**	**0.71**	**0.57–0.89**
Smoking status	No	1		1		1	
	Yes, used to smoke before	**1.52**	**1.21–1.90**	1.23	0.96–1.57	1.15	0.89–1.48
	Yes, currently occasionally	1.4	0.94–2.09	1.37	0.89–2.11	1.2	0.77–1.86
	Yes, currently every day	**1.39**	**1.07–1.82**	0.87	0.65–1.18	0.76	0.55–1.04
Alcohol intake	Never	1		1		1	
	Less than once a month	1.08	0.86–1.36	1.19	0.91–1.55	1.24	0.95–1.63
	Once or a few times a week	**1.61**	**1.16–2.25**	1.36	0.94–1.98	1.37	0.93–2.02
	Once a week	**1.81**	**1.23–2.65**	1.35	0.88–2.07	1.45	0.93–2.26
	(Almost) every day	**2.23**	**1.47–3.41**	1.53	0.96–2.44	**1.63**	**1.00–2.65**
Exercising frequency	Cannot exercise due to injury/illness	1		1		1	
	Once a month or less often	**0.39**	**0.26–0.58**	**0.52**	**0.33–0.81**	**0.59**	**0.37–0.93**
	2–3 times a month	**0.46**	**0.28–0.76**	0.74	0.43–1.27	0.86	0.49–1.50
	Once a week	**0.43**	**0.28–0.67**	0.67	0.42–1.09	0.73	0.45–1.20
	2–3 times a week	**0.35**	**0.23–0.53**	0.6	0.38–0.94	0.65	0.41–1.04
	4–7 times a week	**0.36**	**0.23–0.55**	**0.56**	**0.35–0.90**	**0.6**	**0.37–0.98**
Physical effort required in work	Very little	1		1		1	
	Some	1.18	0.92–1.50	0.93	0.71–1.23	0.97	0.74–1.28
	Average	**1.44**	**1.13–1.83**	0.93	0.70–1.23	0.95	0.71–1.26
	A lot	**1.63**	**1.13–2.35**	0.76	0.50–1.15	0.74	0.49–1.13

Note. The coefficients marked in bold are statistically significant ($p < 0.05$). Model 1: crude ORs for each independent variable separately; Model 2: mutually adjusted for all sociodemographic variables and the respective behavioural variable; Model 3: mutually adjusted for all ten variables.

Table 8

Associations between Socio-demographic and Behavioural Variables and Obesity Status Reported as Odds Ratios with 95% Confidence Intervals

Variable	Category	Model 1		Model 2		Model 3	
		OR	95% CI	OR	95% CI	OR	95% CI
Sociodemographic							
Gender	Male	1		1		1	
	Female	0.87	0.69–1.09	0.85	0.67–1.09	1.07	0.80–1.43
Age group	25-34	1		1		1	
	35-44	**1.87**	**1.26–2.79**	**1.9**	**1.27–2.84**	**1.84**	**1.21–2.79**
	45-54	**3.08**	**2.10–4.51**	**2.91**	**1.96–4.31**	**2.59**	**1.72–3.91**
	55-64	**5.27**	**3.63–7.65**	**5.1**	**3.46–7.50**	**4.35**	**2.88–6.57**
Ethnicity	Estonian	1		1		1	
	Non-Estonian	0.94	0.73–1.22	0.77	0.58–1.02	0.77	0.57–1.03
Education	Primary or basic (1–9 years)	1		1		1	
	Secondary (10–12 years)	0.66	0.44–1.01	**0.58**	**0.37–0.89**	**0.61**	**0.38–0.96**
	Secondary-vocational	0.98	0.68–1.41	0.81	0.55–1.19	0.88	0.59–1.32
	Higher	**0.45**	**0.31–0.66**	**0.46**	**0.31–0.69**	0.52	0.34–0.81
Income	1st quartile	1		1		1	
	2nd quartile	**0.67**	**0.50–0.91**	0.78	0.57–1.07	0.74	0.53–1.03
	3rd quartile	**0.65**	**0.47–0.90**	0.83	0.58–1.18	0.87	0.60–1.25
	4th quartile	**0.48**	**0.34–0.67**	0.71	0.49–1.03	0.73	0.49–1.08
Behavioural							
Chronic disease	Yes	1		1		1	
	No	**0.35**	**0.27–0.44**	**0.45**	**0.35–0.58**	**0.46**	**0.35–0.60**
Smoking status	No	1		1		1	
	Yes, used to smoke before	**1.6**	**1.23–2.09**	**1.49**	**1.12–1.99**	**1.38**	**1.02–1.87**
	Yes, currently occasionally	1.21	0.74–1.97	1.27	0.76–2.12	1.21	0.71–2.06
	Yes, currently every day	1.17	0.84–1.63	0.87	0.61–1.26	0.68	0.46–1.01
Alcohol intake	Never	1		1		1	
	Less than once a month	0.84	0.63–1.13	1.18	0.86–1.64	1.31	0.93–1.84
	Once or a few times a week	1.17	0.80–1.71	1.35	0.88–2.06	**1.37**	**0.88–2.14**
	Once a week	1.44	0.95–2.19	1.54	0.97–2.44	1.69	1.04–2.74
	(Almost) every day	1.49	0.96–2.30	1.53	0.93–2.51	1.58	0.94–2.66
Exercising frequency	Cannot exercise due to injury/illness	1		1		1	
	Once a month or less often	**0.52**	**0.35–0.76**	0.76	0.50–1.15	1.05	0.68–1.62
	2–3 times a month	**0.34**	**0.20–0.58**	0.58	0.32–1.02	0.79	0.43–1.44
	Once a week	**0.45**	**0.29–0.70**	0.72	0.45–1.16	0.94	0.57–1.54
	2–3 times a week	**0.31**	**0.21–0.47**	**0.54**	**0.35–0.84**	0.7	0.44–1.11
	4–7 times a week	**0.28**	**0.18–0.44**	**0.45**	**0.28–0.73**	**0.56**	**0.34–0.92**
Physical effort required in work	Very little	1		1		1	
	Some	0.97	0.71–1.32	**0.71**	**0.51–1.00**	0.77	0.54–1.09
	Average	1.29	0.97–1.71	0.9	0.65–1.25	0.9	0.65–1.26
	A lot	**2**	**1.35–2.95**	1.3	0.84–2.02	1.29	0.82–2.03

Note. The coefficients marked in bold are statistically significant ($p < 0.05$). Model 1: crude ORs for each independent variable separately; Model 2: mutually adjusted for all sociodemographic variables and the respective behavioural variable; Model 3: mutually adjusted for all ten variables.

Table 6 shows the respective results for the associations between obesity and different independent variables. The three regression models were constructed and their results are reported in the same way as in case of overweight, described in the beginning of this section. The results of the models explaining obesity have both similarities with and differences from those explaining overweight. Unlike in case of the latter, where statistically significant associations existed between overweight status and at least one category of all independent variables in Model 1, in case of obesity, the differences between crude odds ratios for gender, ethnicity and for all categories of alcohol intake were not statistically significant. In Model 3, this was also the case with gender, ethnicity, income, and the level of physical effort required in daily work.

Again, the most consistent statistically significant associations in Model 3 could be detected for age group. The respondents in the highest age group (55–64) were 4.35 times (95% CI 2.88–6.57) more likely to be obese than those in the reference group (aged 25–34). Respondents having no chronic disease had considerably lower odds of being obese than those who reported having such a disease (OR 0.46, 95% CI 0.35–0.60). In case of smoking status, only the category of being a former smoker had statistically significant association with obesity – this group was 1.38 (95% CI 1.02–1.87) times more likely to be obese than non-smokers. Interestingly, unlike any other category of the smoking status, this class was statistically significantly associated with obesity in all three models. All categories of education had inverse association with obesity compared to the baseline category (primary or basic education), although the OR estimate was statistically significant only in case of respondents with secondary education (OR 0.61, 95% CI 0.38–0.96). Similarly, in case of alcohol consumption and exercising frequency, only one category of these variables had statistically significant association with obesity compared to the reference category in Model 3 – consuming alcohol once a week (OR 1.37, 95% CI 0.88–2.14) and exercising 4–7 times a week (OR 0.56, 95% CI 0.34–0.92), respectively.

4.3 Predictive modelling

The results in Table 9 show the values of the five metrics (accuracy, precision, recall, F1 and AUC) used for evaluating the performance of the machine learning models compared in this study. The results are shown for the train/test split models, although at the data analysis stage, 10-fold cross-validation was comparatively also applied. In case of predicting overweight, it did not lead to significantly different performance scores from those of train/test split method – the differences between the scores were generally not bigger than a few percentage points. There was more variation in the scores achieved by train/test split and by 10-fold cross-validation method when predicting obesity.

As can be seen from the table, judging by nearly all metrics, all six models performed better when predicting obesity than when predicting overweight status. Of all models, naïve Bayes produced the highest score of AUC in case of predicting both weight categories (0.71 and 0.72, respectively). The performance of the six algorithms is comparatively also shown on figures 7 and 8 that plot the ROC curves of these models. In fact, naïve Bayes emerged as the best performing algorithm by all five metrics, when predicting overweight status but it was less accurate than random forest, logistic regression and decision tree algorithms when predicting obesity (respective figures: 0.67, 0.76, 0.74, and 0.70). In case of the latter, it also had lower sensitivity (recall) and F1 score than these three algorithms. As can also be deducted from the results shown in Table 9, decision tree was the weakest performing model by all metrics when predicting overweight, and it also had the lowest AUC score and

precision when predicting obesity. Logistic regression, in turn, emerged as a middle-ranking or close to top-ranking predictive model by most metrics.

Table 9

Performance Metrics of Six Models when Predicting Overweight and Obesity

Model	Overweight				
	Accuracy	Precision	Recall	F1	AUC
LR	0.63	0.61	0.55	0.58	0.68
DT	0.57	0.53	0.53	0.53	0.57
RF	0.61	0.59	0.53	0.56	0.63
SVM	0.63	0.60	0.57	0.58	0.68
NB	0.66	0.62	0.69	0.65	0.71
KNN	0.62	0.59	0.57	0.58	0.64
	Obesity				
	Accuracy	Precision	Recall	F1	AUC
LR	0.74	0.80	0.89	0.84	0.70
DT	0.70	0.79	0.83	0.81	0.53
RF	0.76	0.81	0.91	0.86	0.65
SVM	0.66	0.88	0.65	0.74	0.71
NB	0.67	0.86	0.69	0.77	0.72
KNN	0.59	0.86	0.57	0.69	0.66

The following two figures (7 and 8) show the ROC curves of all six algorithms.

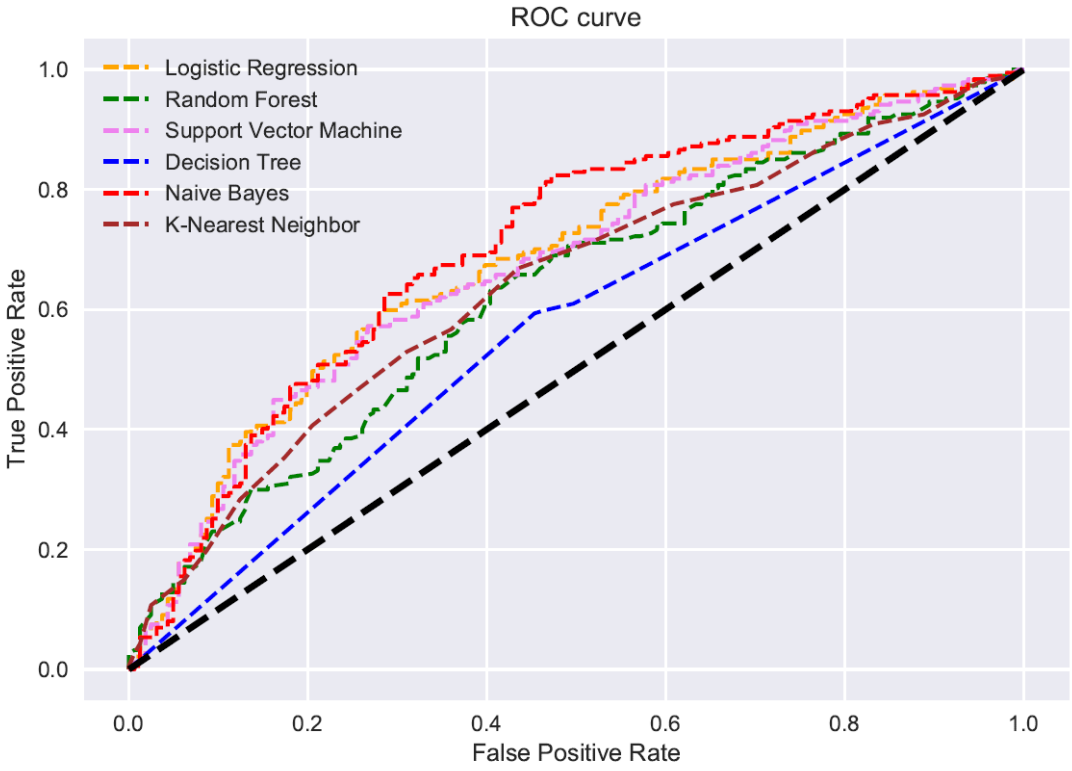


Figure 7

The ROC Curves of Models when Predicting Overweight

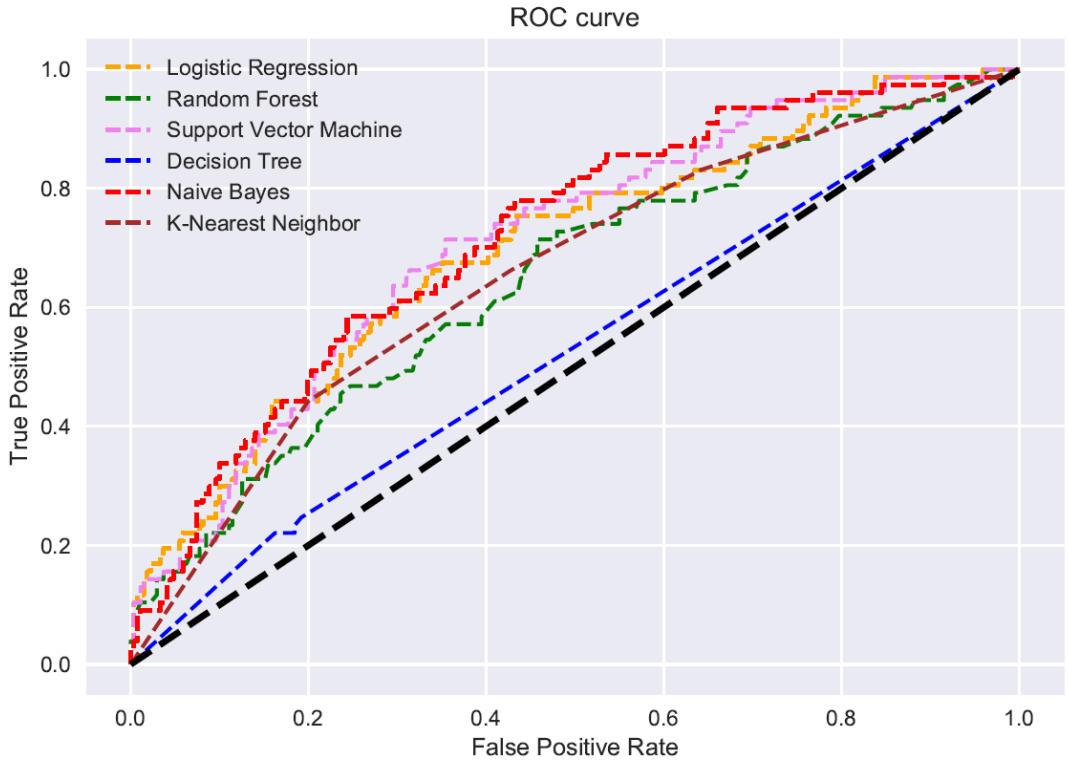
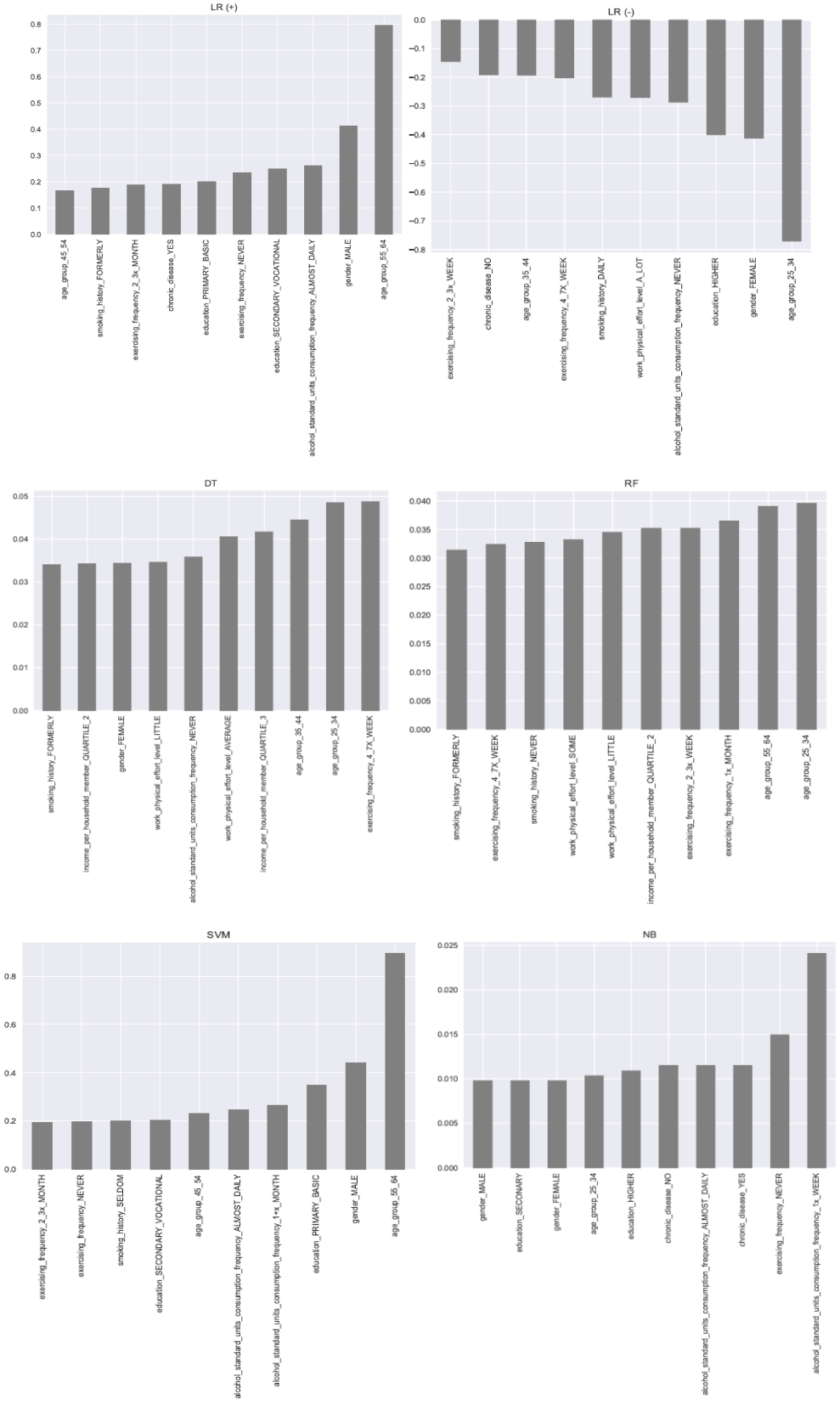


Figure 8
The ROC Curves of Models when Predicting Obesity

Figures 9 and 10, in turn, show the feature importances for each algorithm, except for k-nearest neighbours, because feature importance cannot be calculated directly in case of this particular model. To determine the importance of features, either the attribute *feature_importances_* or *coef_* in Python's sklearn was used, depending on the algorithm. Since no recursive feature elimination was applied in this study, the relevance of all features (in the form of dummy variables to assess the importance of each class of the variable in the model) was compared. For logistic regression, the ten features having the strongest inverse relationship with overweight and obesity, i.e. those with the highest absolute value of negative coefficients are also shown.

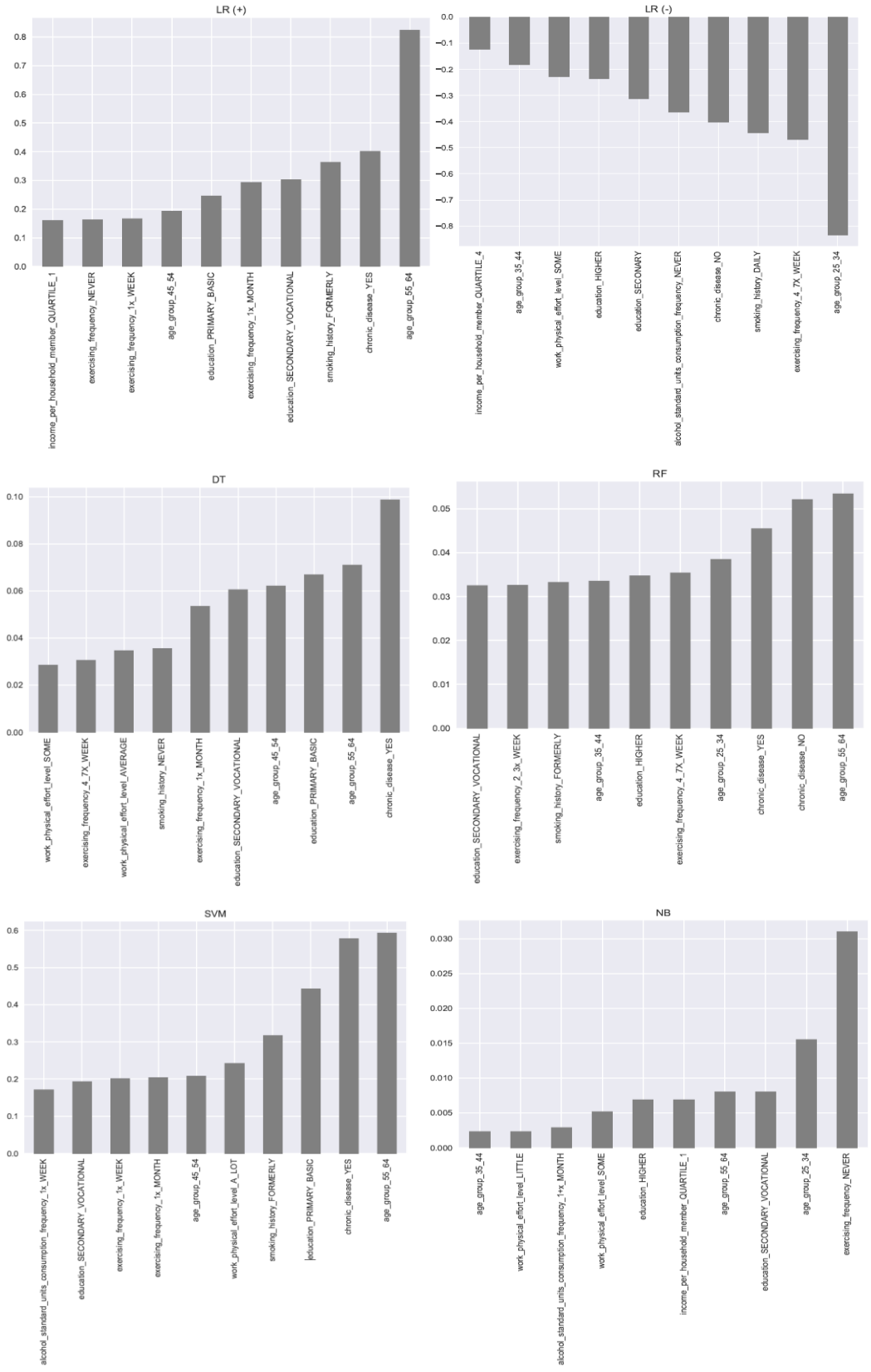
As can be seen from figures 9 and 10, feature importance varied across different algorithms which is not surprising since their underlying computational mechanisms differ. Feature importance also varied depending on whether the algorithm was used for predicting overweight or obesity. When predicting overweight, one single feature was significantly more important than all others in case of logistic regression, support vector machine, and naïve Bayes – belonging to the highest age-group (55–64) in case of the first two and exercising once a week in case of the latter. In case of decision tree and random forest algorithms, the differences between the importance of top features in terms of contributing to the model performance were marginal. When predicting obesity, however, the importance of one single feature was clearly distinguishable from that of others in case of logistic regression, decision tree, and naïve Bayes algorithms. Also, in three models out five, the most significant feature of the model was the respondent's belonging to the highest age-group (55–64). In case of decision tree, however, the existence of chronic disease, and in case of naïve Bayes, belonging to the category of not exercising at all contributed the most to the predictive performance of the model.



Note. Ten most important features contributing to prediction of overweight are shown for decision tree (DT), random forest (RF), support vector machine (SVM), and naïve Bayes (NB). In case of logistic regression (LR), ten features with highest absolute values for both positive (LR(+)) and negative (LR(-)) coefficients are shown. In case of KNN algorithm, feature importances cannot be directly calculated.

Figure 9

Feature Importance in Models (Prediction of Overweight)



Note. Ten most important features contributing to prediction of obesity are shown for decision tree (DT), random forest (RF), support vector machine (SVM), and naïve Bayes (NB). In case of logistic regression (LR), ten features with highest absolute values for both positive (LR(+)) and negative (LR(-)) coefficients are shown. In case of KNN algorithm, feature importances cannot be directly calculated.

Figure 10

Feature Importance in Models (Prediction of Obesity)

5 Discussion

The discussion chapter is divided into three sections, first focusing separately on the results of explanatory and predictive modelling and then at a higher level on the comparison of and a possible synergy between the two approaches.

5.1 Explanatory modelling

The outcomes of regression analysis were in essence not unexpected – where statistically significant associations between socio-demographic or behavioural factors and excess body weight were detected, these were consistent, by and large, with the results of various other studies on overweight and obesity using explanatory modelling. However, some interesting disagreements with other studies also emerged and various outcomes are worthy of some further and more in-depth scrutiny.

In models simultaneously adjusted for all ten independent variables, overweight had statistically significant association with the dichotomous variables of gender, ethnicity, and chronic disease, as well as with all classes of age group. Additionally, there was statistically significant association between being overweight and distinct classes of some other ordinal variables, namely education, alcohol intake, and exercising frequency. Interestingly, however, gender and ethnicity had no statistically significant association with obesity. This outcome is in disagreement with the results of various other studies that have either singled out Estonian men in particular in terms of obesity (Webber et al. 2010) or established a more general relationship between gender and obesity in Estonia (Reile et al. 2020; Reile and Leinsalu 2019) or elsewhere (Seo and Li 2009). Similarly to being overweight, a graded increase of obesity level was detected in case of age group. This concurs with the results of various other studies that have established a strong association between age and obesity in Estonia (Marques et al. 2017; Reile and Leinsalu 2019; Tekkel, Veideman, and Rahu 2010;). Like in case of overweight, obesity had statistically significant association with some categories of education and alcohol intake, and – contrary to overweight – also with smoking status.

Slightly surprisingly, regression analysis did not reveal a statistically significant association between income and excess body weight – neither in case of overweight, nor obesity. This contradicts the results reported in studies in various other contexts that have demonstrated a strong relationship between socioeconomic factors and obesity in particular (Marques et al. 2017; Paeratakul et al. 2002), although the relationship has not been claimed to be an unequivocally inverse one (Clarke et al. 2008).

In case of education, the fact that the odds of being overweight or obese were lower for nearly all education levels compared to the baseline category of basic or primary education is consistent with the results of various other studies (Clarke et al. 2008; Tekkel, Veideman, and Rahu 2010; Maher et al. 2013), although the association was statistically significant only for selected categories.

The association between smoking status and overweight and obesity also merits particular scrutiny here. The fact that the odds of being overweight or obese were the highest for former smokers, although statistically significant only in case of obesity, is interesting. Nuttall (2015, 123), based on the data from the United States, claims that the rise of obesity and the dramatic decrease in smoking in last decades might be linked, and that the latter can actually be one of the contributing factors to the "obesity epidemic" and the fact that not only has the mean BMI augmented, but that there has also been a growing increase in skewing of the BMI

distribution in the US population toward the right, i.e. toward very large BMI values. This is probably more appropriate an explanation for the trends in the US context, but in general terms Nuttall's argumentation could also explain the findings of the present study. As Nuttall (*ibid*.) suggests, smoking contributes to a lower BMI by at least two ways: it impairs appetite and it can also lead to the development of chronic obstructive pulmonary disease, which results in a lower body mass. Quitting smoking can, conversely, lead to the rise of the BMI which explains why former smokers from among all the categories of the smoking status feature are the most likely ones to be obese.

Another result worth following up here more in detail is the fact that higher levels of exercising were inversely associated with overweight and obesity. The most frequent exercisers (4–7 times a week) had considerably lower odds of being overweight or obese at statistically significant level compared to the individuals who did not exercise at all. These odds were lower, in fact, for all exercising categories compared to the reference category although the decreasing trend of odds while exercising level increased was not consistent, probably owing to the fact that a rather complex scale was used in the questionnaire. Also, the self-reported data are not necessarily reliable. The inverse association between exercising levels and BMI has been demonstrated by numerous studies also mentioned in section 2.3 (Maher et al. 2013; Seo and Li 2009; Stamatakis et al. 2008; Wanner et al. 2016), as well as various others (Besson et al. 2009; Cheng et al. 2021; Wanner et al. 2017). It is important to note, however, that an unequivocal causal relationship between exercising levels and excess body weight cannot be established, especially in case of a study that follows a cross-sectional design. Indeed, it is generally assumed, that the level of physical activity is a determinant of the amount of body fat due to energy expenditure (Chin et al. 2016; Wanner et al. 2017; Wareham 2005) but a possibility of reverse causality cannot be ruled out. This means that it could be the amount of excess body weight that actually determines the level of physical activity, and that people who are overweight or obese exercise less namely due to that fact (Wanner et al. 2017, 192). The possibility of similar reverse causality has also been argued in case of the relationship between the BMI and the amount of sedentary time that could be interpreted as yet another manifestation of physical activity or the lack thereof (Ekelund et al. 2008; Pedisic et al. 2014; Pulsford et al. 2013). Pulsford et al. (2013), for example, conclude in their study that sitting time is not associated with obesity cross-sectionally or prospectively. Ekelund et al. (2008) conclude, in turn, that while BMI might predict sedentary time, the latter does not predict future obesity.[13]

And finally, the fact that the likelihood of being overweight or obese grew with the increased level of alcohol intake, although the association was statistically significant only in case of selected levels, could be explained at least partly by hidden (although not detected) multi-collinearity between this and certain sociodemographic factors such as the level of education, as well as adverse behavioural characteristics such as low level of physical activity or being a smoker. The same applies to the higher likelihood of being obese among those respondents whose work required considerable level of physical effort, as these jobs are generally also associated with lower level of education and income.

5.2 Predictive modelling

As the results of the evaluation and comparison of six machine learning algorithms commonly used for solving classification problems revealed, their performance varied,

[13] Incidentally, in the present study a similar kind of reverse causality can potentially also exist in case of the association between excess body weight and having a chronic disease.

although in general not very significantly. The performance of none of the algorithms was outstanding, although in all cases except for the decision tree their predictions were better than would have been achieved by random classification (i.e. in most cases well above 50 percent). In case of predicting overweight, naïve Bayes emerged as the best performing model by all evaluation metrics scrutinised, although according to none of the metrics was its performance significantly better than that of at least some other models. Its AUC score (0.71), for example, was only marginally higher than that of logistic regression and support vector machine models. In case of predicting obesity, no one single model emerged as having unequivocally best performance by all metrics. For example, naïve Bayes had again the highest AUC score (0.72) but ranked fourth in terms of overall accuracy.

These figures and the comparison between different models should be considered very circumstantial though – as the aim was to compare the models in general terms and to employ the same set of features that was used in case of explanatory modelling, mostly default hyperparameters were used when training the models, further calibration and optimisation of hyperparameters was not applied, and no recursive elimination of features was performed. The fine-tuning of the models might have led to their better performance as well as changed their ranking by the used evaluation metrics. It is important to note that in the current methodological setting, 10-fold cross-validation did not lead to significantly different performance scores of these models when predicting overweight compared to the train/test split method. In case of predicting obesity, some metrics were more substantially affected by using 10-fold cross-validation but without further model tuning and optimisation the significance of these differences is difficult to assess.

A consistent trend worth highlighting here, however, was a better performance of most models by nearly all metrics when predicting obesity, compared to the prediction of overweight. This was particularly manifested in case of precision and recall, and consequently also the more robust F1-score. This means that when predicting obesity, the models were relatively successful when identifying positive cases of obesity from all true positives, and the number of false positives predicted by the models was relatively lower than in case of predicting overweight. Such performative difference between predicting overweight and obesity can owe to the fact that obesity as a weight category is located at the extreme end of the BMI spectrum and is thus more likely to have a strong association with certain socio-demographic and especially behavioural factors. The category "overweight" in turn, was in this study referring to a rather broad and hence also a heterogeneous section of the BMI spectrum, i.e. to anyone with BMI $\geq$ 25.0 kg/m^2. Additionally, the lower end of the overweight category (i.e. the BMI values relatively close to 25.0 kg/m^2) could be considered a particularly "grey area" of the BMI spectrum, where the limitations of the metric as a reliable indicator outlined in section 2.1 are especially prominent. The persons categorised as "slightly overweight" based on the value of their BMI might be such by virtue of their idiosyncratic body composition and type, and not, for example, because of low level of exercising or other adverse behavioural factors.

The position of logistic regression in the comparison of the performance of predictive models also deserves special scrutiny. As the results of the comparison show, logistic regression was roughly middle- or close to top-ranking by all metrics in case of predicting both overweight and obesity. Its AUC score, accuracy, and precision were not very much below those of naïve Bayes, the best performing model in case of overweight prediction. The same can be concluded about all evaluation metrics of the logistic regression model when predicting obesity – the scores were only slightly lower than those of the best performing model in the respective category.

How do these results compare with those reported in other studies that have used predictive modelling when scrutinising overweight and obesity? In terms of pure metrics, the values achieved in this study are middle-ranking compared to the ones achieved by other studies. Ferdowsy et al. (2021), for example, report 97 percent accuracy of logistic regression when predicting obesity in Bangladesh, Dugan et al. (2015) report 85 percent accuracy of naïve Bayes when predicting early childhood obesity in the United States, and Cañas and Martinez (2020) report the value as high as 98 percent for AUC when predicting obesity among students in three Latin American countries. Conversely, Kim et al. (2019) reported accuracy values slightly above and below 50 percent for all ten classifiers when predicting obesity among adolescents in South Korea, and AUC scores ranging between 67 and 76 percent. Pang et al. (2021), in turn, reported the accuracy values ranging between 62 and 66 percent when predicting obesity with seven algorithms among children aged 2–7 in the United States.

Comparison with other studies makes more sense at a high level though, since the context of the studies, datasets sizes, and input feature sets differ considerably, training and modelling techniques likewise. Obviously, the results also depend on the concrete sociodemographic group under scrutiny, since the reasons for excess body weight differ for children, adolescents, and adults. Rather than comparing the actual values of algorithm performance, it is thus more illuminating to contextualise the results of this study in broader terms. As was already demonstrated in section 2.4 and Table 2, all nine studies described unequivocally singled out one algorithm that outperformed all others, in some cases significantly (Cañas and Martinez 2020; Ferdowsy et al. 2021) but in most cases just by a narrow margin (Cheng et al. 2021; Curbelo et al. 2017; Delnevo et al. 2021; Kim et al. 2019; Pang et al. 2021; Thamrin et al. 2021). For example, when comparing the performance of 11 algorithms to predict overweight and obesity among the US population, Cheng et al. (2019) reported the accuracy of nearly all of them within a narrow range of 1.5 percent, and the sensitivities within a range of 3 percent. In two of these studies, logistic regression emerged as the best performing algorithm (Ferdowsy et al. 2021; Thamrin et al.), while in others that included logistic regression among compared models, its performance was middle-ranking in terms of most metrics scrutinised (Cheng et al. 2019; Kim et al.; Pang et al. 2021).

It can be thus argued that in general terms the results of this thesis are in agreement with the conclusions of those studies that have demonstrated a slight but not necessarily overwhelmingly better performance of one particular model compared to others and placed logistic regression as an estimator roughly in the middle or close to the top of the pack in terms of its predictive performance.

5.3 Reconciling the two "modelling cultures"

And finally, it is pertinent to ponder comparatively and at a higher level upon the advantages and limitations of explanatory and predictive modelling in the research on overweight and obesity, based on the outcomes of this study. Although explanatory and predictive models serve different purposes, their methods are partly overlapping, as was already argued in Chapter 2. The "bridge" between the two modelling cultures in this study has been logistic regression – a conventional statistical method for scrutinising the potential associations between dependent and independent variables but also a classificatory tool that can be trained like more advanced machine learning models for predicting the output from the input data. There are two different angles from which logistic regression can be compared with other methods used in this study – namely, feature importance and predictive performance.

As was shown in section 4.3, the ten most important features for predicting overweight and obesity differed considerably depending on the model under scrutiny which can be explained by the differences in how these models computationally operate. This, however, means that the results of feature importance analysis in case of machine learning models – also in this study – are often intuitively difficult if not impossible to interpret. Logistic regression could be considered an exception in this regard. The higher the absolute value of the positive or negative coefficient of the feature, the more it contributes to or inhibits the probability of being overweight or obese. Also, the features importances calculated for logistic regression as a predictive model in this study were compatible with the results of regression analysis. No definite conclusions about the advantages or disadvantages of different modelling approaches can be drawn from this, although a closer look at feature importances in models does tend to point to a common argument that more advanced machine learning algorithms might have a good predictive capacity but produce a "black box" model that is not generally helpful for explaining the possible causal links between variables.

Moreover, the arguments for an outstandingly better performance of more advanced predictive models compared to logistic regression, and, more generally, that of machine learning as such compared to conventional statistical modelling, are not supported by the results of this study. Interestingly, the position of different authors with regard to the comparative performance of standard regression and machine learning models differs diametrically. Breiman's (2001) and Hindman's (2015) clear preference for the latter was already highlighted in section 2.2. In fact, Hindman (*ibid*.) vocally advocates replacing typical regression analyses, in social sciences especially, with multi-algorithm ensemble approaches. More specifically in case of studying overweight and obesity and while comparing different modelling approaches, Delnevo et (2021, 11) conclude that machine learning outperforms traditional statistics. Colmenarejo (2020, 24), in an overview article of many studies of overweight and obesity, claims that when in the same study logistic or linear regression has been compared with machine learning models, the latter have given better results in terms of prediction performance and accuracy. However, claims opposite to those by Colmenarejo and other proponents of machine learning have also been made, showing in various contexts that machine learning models do not necessarily outperform standard regression (Cheng et al. 2021; Christodoulou et al. 2019; Doupe et al. 2019). Christodoulou et al. (2019), for example, analysed the results of 71 articles that compared the performance of logistic regression with that of advanced machine learning and found no evidence of the superiority of the latter. In fact, as various studies also reviewed in section 2.4 have demonstrated, logistic regression can outperform more advanced models (Ferdowsy et al. 2021; Thamrin et al. 2021). As Cheng et al. (2021, 9) conclude while reviewing many research accounts, advanced machine learning methods with high computational complexity might thus not always be necessary for predicting obesity and owing to its well-understood theoretical and computational background, the logistic regression models could be "enough."

The above discrepancies of opinion further highlight the argument already presented in section 2.2 – explanatory and predictive modelling form a complex continuum and their juxtaposition or unconditional preference for one over the other is not necessarily justified and useful, at least not in case of studying overweight and obesity.

6 Conclusions

The aims of this thesis were threefold. Based on the 2020 Health Behaviour Among Estonian Adult Population survey data it first sought to scrutinise overweight and obesity from the perspective of explanatory modelling. Using the conventional method of (binary) logistic regression, it inspected the possible associations between being overweight or obese and ten different socio-demographic and behavioural variables. Secondly, the study approached the same data from the perspective of predictive modelling, by comparing the performance of six machine learning algorithms commonly used for classification problems, including logistic regression, when predicting overweight and obesity status. And finally, based on the results of the two modelling approaches, the study aimed to engage selectively with other research that has been done on these topics, as well as to offer a higher-level discussion of the advantages and limitations of explanatory and predictive modelling. The conclusions of the study, in a nutshell, are as follows.

Firstly, the regression analysis model mutually adjusted for all ten independent variables revealed that overweight and obesity had statistically significant association with at least one category of all independent variables, except for the physical effort required by daily work. Age and having a chronic disease had the most consistent association with both dependent variables. The fact that gender, ethnicity, and at least some categories of education, income, smoking status, alcohol intake, and exercising frequency had statistically significant association with either overweight or obesity, is not necessarily surprising when compared to the results of other explanatory studies done on the topic but contributes nevertheless to a better understanding of the factors influencing overweight and obesity prevalence in Estonia.

Secondly, the results of predictive modelling demonstrated a performance that was better than random classification in case of nearly all machine learning algorithms, although their predictions were generally more accurate for obesity than for overweight. Naïve Bayes was the best performing algorithm, although not overwhelmingly so, by all metrics in case of predicting overweight. It had also the highest AUC score when predicting obesity, while the random forest algorithm had the highest accuracy, recall, and F1-score. Logistic regression was middle- or close to top-ranking algorithm by nearly all metrics in both overweight and obesity prediction. In broad terms, the results of the comparison of machine learning models based on the Estonian data were in accordance with those reported in at least some other studies discussed comparatively in this thesis. However, a considerable lack of unanimity can be detected in the research accounts that have applied predictive modelling in the study overweight and obesity – conclusions about the best-performing algorithms and even about the feasibility of the whole approach differ radically.

And finally, the results of this study do not provide strong evidence for preferring predictive modelling over explanatory one or vice versa. Instead of an outright juxtaposition of the two "modelling cultures" that has often led the respective authors to argue for the advantages of machine learning over "traditional" statistics, especially in the context of dealing with increasingly big and also wide data (Breiman 2001; Hindman 2015), an approach that considers the two complementary to each other can be more appropriate. The author of this thesis agrees with Colmenarejo (2020, 26) that the choice between the two is not only a matter of the size of the data but it also depends on what are the research aims. If these are interpretability, inference, and theoretical clarity based on relatively simple models with relatively few predictor variables, explanatory modelling based on classical statistics serves these purposes generally better (*ibid*.). If, however, the purpose is a very good predictive performance and interpretability or inference are not among the priorities of the study, predictive modelling by means of machine learning methods can be prioritised (*ibid*.).

Like any other study, this one too has various limitations that need to be acknowledged. Firstly, despite the wide acceptance of BMI as a metric, it has various limitations that were already highlighted in section 3.1. Secondly, all data in the Health Behaviour Among Estonian Adult Population survey are self-reported and not objectively measured – the answers to the survey questions can thus be biased. This is particularly relevant in case of self-reported height and weight that were used to calculate the BMI values and determine the corresponding BMI categories that were used as dependent variables in this thesis. As some studies have shown (eg. Bowman and DeLucia; Clarke 2000; Oja et al. 2020), the figures for self-reported height tend to be higher and those for self-reported weight lower than the objectively measured ones, thus not allowing for the calculation of true BMI values. Moreover, such misreporting can be affected by socio-demographic factors (Aasvee et al. 2015; Nawaz et al. 2001; Nyholm 2007). Thirdly, the independent variables in this study did not cover data on dietary behaviour, owing to the fact that questions on nutrition and eating in the Health Behaviour Among Estonian Adult Population survey generally refer to the food intake during the week prior to answering the question and are thus not reliable indicators of long-term eating habits. Fourthly, from the perspective of explanatory modelling in particular, the cross-sectional study design of the survey does not allow to provide strong evidence of the existence of causal relationship between overweight and obesity, and the independent variables scrutinised. Nor can reverse causality as an alternative explanation of some detected associations between dependent and independent variables be ruled out. And fifthly, for the purposes of predictive modelling, the dataset was relatively small and a larger dataset might have been beneficial for model training and hence potentially improved the performance of some or all models.

The results as well as the limitations of this study provide various clues pointing to possible directions for further research. In case of regression analysis, thought could be given to using an alternative set of independent variables and stepwise regression analysis could be used to eliminate the features from the eventual model that do not improve model performance. It would also be interesting to examine cross-sectionally various more long-term trends of overweight and obesity in Estonia, rather than focusing only on one survey year as was done in this thesis. Also, the BMI category itself could be employed as an independent variable instead of being a dependent one – for example in the study of factors affecting physical activity levels and exercising frequency. In case of predictive modelling, this study did not aim at determining the best-tuned versions of the compared algorithms when predicting overweight or obesity. In pragmatic research with clinical or policy-making aims where the accuracy of the prediction is a priority, however, a more rigorous approach to model-training that involves advanced validation techniques as well as further calibration and optimisation of hyperparameters could be used, possibly while comparing an even broader set of predictive models.

7 References

Aasvee, K., Rasmussen, M., Kelly, C., Kurvinen, E., Giacchi, M., and Ahluwalia N. 2015. Validity of Self-Reported Height and Weight for Estimating Prevalence of Overweight Among Estonian Adolescents: The Health Behaviour in School-Aged Children Study. *BMC Research Notes* 8(606). DOI: https://doi.org/10.1186/s13104- 015-1587-9

Adnan, M.H.M., Husain, W., and Rashid, N.A. 2012. Hybrid Approaches Using Decision Tree, Naïve Bayes, Means and Euclidean Distances for Childhood Obesity Prediction. *The International Journal of Software Engineering & Applications* 6(3), 99-106. DOI: https://doi.org/10.1109/ICCISci.2012.6297254

Afshin, A., Forouzanfar, M.H., Reitsma, M.B., et al. 2017. Health Effects of Overweight and Obesity in 195 Countries Over 25 Years. *The New England Journal of Medicine* 377(1), 13−27. DOI: https://doi.org/10.1056/NEJMoa1614362

Ahmadi, M. N., Pavey, T. G., and Trost, S. G. 2020. Machine Learning Models for Classifying Physical Activity in Free-Living Preschool Children. *Sensors* 20(16), 4364. DOI: https://doi.org/10.3390/s20164364

Ahmed, M.U., and Loutfi, A. 2013. Physical Activity Identification Using Supervised Machine Learning and Based on Pulse Rate. *International Journal of Advanced Computer Science and Applications* 4(7), 210–217. DOI: http://dx.doi.org/10.14569/IJACSA.2013.040730

Besson, H., Ekelund, U., Luan, J., May, A.M., Sharp, S., Travier, N. et al. 2009. A Cross-Sectional Analysis of Physical Activity and Obesity Indicators in European Participants of the EPIC-PANACEA Study. *International Journal of Obesity* 33, 497–506. DOI: https://doi.org/10.1038/ijo.2009.25

Bixby, H., Bentham, J., Zhou, B., et al. 2019. Rising Rural Body-Mass Index Is the Main Driver of the Global Obesity Epidemic in Adults. *Nature* 569(7755), 260−281. DOI: https://doi.org/10.1038/s41586-019-1171-x

Bowman, R.L. and DeLucia J.L. 1992. Accuracy of Self-reported Weight: A Meta-analysis. *Behavor Therapy* 23, 637–655. DOI: https://doi.org/10.1016/S0005-7894(05)80226-6

Breiman, L. 2001. Statistical Modeling: The Two Cultures. *Statistical Science* 16(3), 199–231. DOI: https://doi.org/10.1214/ss/1009213726

Buskirk, T.D., Kirchner, A., Eck, A. and Signorino, C.S. 2018. An Introduction to Machine Learning Methods for Survey Researchers. *Survey Practice* 11(1). DOI: https://doi.org/10.29115/SP-2018-0004

Bzdok, D., Altman, N., and Krzywinski, M. 2018. Statistics versus Machine Learning. *Nature Methods* 15(4), 233–234. DOI: https://doi.org/10.1038/nmeth.4642

Caballero, B. 2007. The Global Epidemic of Obesity: An Overview. *Epidemiologic Reviews* 29(1), 1–5. DOI: https://doi.org/10.1093/epirev/mxm012

Cañas Cervantes, Rodolfo Ubaldo Martinez Palacio 2020. Estimation of Obesity Levels Based on Computational Intelligence. *Informatics in Medicine Unlocked* 21, 100472. DOI: https://doi.org/10.1016/j.imu.2020.100472

Chatterjee, A., Gerdes, M.W., and Martinez, S.G. 2020. Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview. *Sensors* 20(9), 2734. DOI: https://doi.org/10.3390/s20092734

Cheng, X., Lin, S., Liu, J., et al 2021. Does Physical Activity Predict Obesity—A Machine Learning and Statistical Method-Based Analysis. *International Journal of Environmental Research and Public Health* 18, 3966. DOI: https://doi.org/10.3390/ijerph18083966

Chin, S.-H., Kahathuduwa, C.N., and Binks, M. 2016. Physical Activity and Obesity: What We Know and What We Need to Know. *Obesity Reviews* 17, 1226–1244. DOI: https://doi.org/10.1111/obr.12460

Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., and Van Calster, B. 2019. A Systematic Review Shows no Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *The Journal of Clinical Epidemiology* 110, 12–22. DOI: https://doi.org/10.1016/j.jclinepi.2019.02.004

Clarke P, O'Malley, P.M., Johnston, L.D., and Schulenberg, J.E. 2009. Social Disparities in BMI Trajectories Across Adulthood by Gender, Race/ Ethnicity and Lifetime Socio-Economic Position: 1986–2004. *The International Journal of Epidemiology* 38(2), 499–509. DOI: https://doi.org/10.1093/ije/dyn214

Colmenarejo, G. 2020. Machine Learning Models to Predict Childhood and Adolescent Obesity: A Review. *Nutrients* 12(8), 2466. DOI: https://doi.org/10.3390/nu12082466

Courtemanche, C.J., Pinkston, J.C., Ruhm, C.J., and Wehby, G.L. 2016. Can Changing Economic Factors Explain the Rise in Obesity? *Southern Economic Journal* 82, 1266–1310. DOI: https://doi.org/10.1002/soej.12130

Csige, I., Ujvárosy, D., Szabó, Z., Lorincz, I., Paragh, G., Harangi, M., and Somodi, S. 2018. The Impact of Obesity on the Cardiovascular System. *Journal of Diabetes Research* 2018, 3407306. DOI: https://doi.org/10.1155/2018/3407306

Curbelo Montañez, Fergus, P., Hussain, A., et al. 2017. Machine Learning Approaches for the Prediction of Obesity Using Publicly Available Genetic Profiles. *2017 International Joint Conference on Neural Networks (IJCNN)*, 2743-2750. DOI: https://doi.org/10.1109/IJCNN.2017.7966194

DeGregory, K.W., Kuiper, P., DeSilvio, T., et al. 2018. A Review of Machine Learning in Obesity. *Obesity Reviews* 19, 668–685. DOI: https://doi.org/10.1111/obr.12667

Delnevo, G., Mancini, G., Roccetti, M., Salomoni, P., Trombini, E., and Andrei, F. 2021. The Prediction of Body Mass Index from Negative Affectivity through Machine Learning: A Confirmatory Study. *Sensors* 21(), 2361. DOI: https://doi.org/10.3390/s21072361

Doupe, P., Faghmous, J., and Basu, S. 2019. Machine Learning for Health Services Researchers. *Value Health* 22, 808–815.DOI: https://doi.org/10.1016/j.jval.2019.02.012

Dugan, T.M., Mukhopadhyay, S., Carroll, A., and Downs, S. 2015. Machine Learning Techniques for Prediction of Early Childhood Obesity. *Applied Clinical Informatics* 6(3), 506–520. DOI: https://doi.org/10.4338/ACI-2015-03-RA-0036

Dunstan, J., Aguirre, M., Bastías, M., et al. 2020. Predicting Nationwide Obesity from Food Sales Using Machine Learning. *Health Informatics Journal* 26, 652–663. DOI: https://doi.org/10.1177/1460458219845959

Ekelund, U., Brage, S., Besson, H., Sharp, S., and Wareham, N.J. 2008. Time Spent Being Sedentary and Weight Gain in Healthy Adults: Reverse or Bidirectional Causality? *The American Journal of Clinical Nutrition* 88(3), 612–617. DOI: https://doi.org/10.1093/ajcn/88.3.612

Eurostat 2021. Over Half of Adults in the EU Are Overweight. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20210721-2 (accessed on 27 September 2021)

Ezzati, M., Bentham, J., Di Cesare, M., et al. 2017. Worldwide Trends in Body-Mass Index, Underweight, Overweight, and Obesity from 1975 to 2016: A Pooled Analysis of 2416 Population-Based Measurement Studies in 128.9 Million Children, Adolescents, and Adults. *Lancet* 390(10113), 2627−2642. DOI: https://doi.org/10.1016/S0140-6736(17)32129-3

Ferdowsy, F., Rahi, K., Jabiullah, I., and Habib, T. 2021. A Machine Learning Approach for Obesity Risk Prediction. *Current Research in Behavioral Sciences* 2, 100053. DOI: https://doi.org/10.1016/j.crbeha.2021.100053.

Fogelholm, M and Kukkonen-Harjula K. 2000. Does Physical Activity Prevent Weight Gain–A Systematic Review. *Obesity Reviews* 1(2), 95–111. DOI: https://doi.org/10.1046/j.1467-789x.2000.00016.x

Hindman, M. 2015. Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science* 659(1), 48–62. DOI: https://doi.org/10.1177/0002716215570279

Jindal, K., Baliyan, N., and Rana, P.S. 2018. Obesity Prediction Using Ensemble Machine Learning Approaches. In Sa P., Bakshi S., Hatzilygeroudis I., Sahoo M. (eds) *Recent Findings in Intelligent Computing Techniques. Advances in Intelligent Systems and Computing*, vol. 708. Springer, Singapore, 355–362. DOI: https://doi.org/10.1007/978-981-10-8636-6_37

Johnson, W., Li, L., Kuh, D., and Hardy R. 2015. How Has the Age-Related Process of Overweight or Obesity Development Changed Over Time? Co-Ordinated Analyses of Individual Participant Data from Five United Kingdom Birth Cohorts. *PLOS Medicine* 12, 1001828. DOI: https://doi.org/10.1371/journal.pmed.1001828

Jordan, M.I., and Mitchell, T.M. 2015. Machine Learning: Trends, Perspectives, and Prospects. *Science* 349 (6245), 255-60. DOI: https://www.science.org/doi/10.1126/science.aaa8415

Katus, U., Villa, I., Ringmets, I. et al. 2020. Association of FTO rs1421085 with Obesity, Diet, Physical Activity and Socioeconomic Status: A Longitudinal Birth Cohort Study. *Nutrition, Metabolism & Cardiovascular Diseases* 30(6), 948–959. DOI: https://doi.org/10.1016/j.numecd.2020.02.008

Kern, C., Klausch, T., and Kreuter F. 2019. Tree-based Machine Learning Methods for Survey Research. *Survey Research Methods* 13(1), 73–93. DOI: https://doi.org/10.18148/srm/2019.v1i1.7395

Kim, C., Costello, F.C., Change, K., et al. 2019. Predicting Factors Affecting Adolescent Obesity Using General Bayesian Network and What-If Analysis. *International Journal of Environmental Research and Public Health* 16(23), 4684. DOI: https://doi.org/10.3390/ijerph16234684

Kirby, J.B., Liang, L., Chen, H.-J., and Wang, Y. 2012. Race, Place, and Obesity: The Complex Relationships Among Community Racial/Ethnic Composition, Individual Race/Ethnicity, and Obesity in the United States. *The American Journal of Public Health* 102, 1572–1578. DOI: https://doi.org/10.2105/AJPH.2011.300452

Maher, C.A., Mire, E., Harrington, D.M., Staiano, A.E., and Katzmarzyk, P.T. 2013. The Independent and Combined Associations of Physical Activity and Sedentary Behavior with Obesity in Adults: NHANES 2003-06. *Obesity* 21(12), 730–737. DOI: https://onlinelibrary.wiley.com/doi/10.1002/oby.20430

Marques, A., Peralta, M., Naia, N., Loureiro, N., and Gaspar de Matos, M. 2018. Prevalence of Adult Overweight and Obesity in 20 European Countries, 2014. *The European Journal of Public Health* 28(2), 295–300. DOI: https://academic.oup.com/eurpub/article/28/2/295/4210290

Molina, M. and Garip, F. 2019. Machine Learning for Sociology. *Annual Review of Sociology* 45, 27–45. DOI: https://doi.org/10.1146/annurev-soc-073117-041106

Nawaz, H., Chan, W., Abdulrahman, M.; Larson, D., and Katz, D.L. 2001. Self-Reported Weight and Height: Implications for Obesity Research. *The American Journal of Preventive Medicine* 20(4), 294–298. DOI: https://doi.org/10.1016/S0749-3797(01)00293-8

Ngiam, K.Y. and Khor, I.W. 2019. Big Data and Machine Learning Algorithms for Health-Care Delivery. *The Lancet Oncology* 20(5), 262–273. DOI: https://doi.org/10.1016/S1470-2045(19)30149-4

Nuttall, F.Q. 2015. Body Mass Index: Obesity, BMI, and Health: A Critical Review. *Nutrition Today* 50(3), 117–128. DOI: https://doi.org/10.1097/NT.0000000000000092

Nyholm, M., Gullberg, B., Merlo, J. et al. The Validity of Obesity Based on Self-Reported Weight and Height: Implications for Population Studies. *Obesity* 15(1), 197–197. DOI: https://doi.org/10.1038/oby.2007.536

Oja, L., Slapšinskaitė, A., Piksööt, J., and Šmigelskas, K. 2020. Baltic Adolescents' Health Behaviour: An International Comparison. *International Journal of Environmental Research and Public Health* 17(22), 8609. DOI: https://doi.org/10.3390/ijerph17228609

Paeratakul, S., Lovejoy, J., Ryan, D., and Bray, G. 2002. The Relation of Gender, Race and Socioeconomic Status to Obesity and Obesity Comorbidities in a Sample of US Adults. *International Journal of Obesity* 26, 1205–1210. DOI: https://doi.org/10.1038/sj.ijo.0802026

Pang, X., Forrest, C.B., Lê-Scherban, F., and Masino, A.J. 2021. Prediction of Early Childhood Obesity with Machine Learning and Electronic Health Record Data. *International Journal of Medical Informatics* 150, 104454. DOI: https://doi.org/10.1016/j.ijmedinf.2021.104454

Park, B., Chung, C.-S., Lee, M.J., and Park, H. 2020. Accurate Neuroimaging Biomarkers to Predict Body Mass Index in Adolescents: A Longitudinal Study. *Brain Imaging and Behavior* 14, 1682-1695. DOI: https://doi.org/10.1007/s11682-019-00101-y

Pedisic, Z, Grunseit, A., Ding, D et al. 2014. High Sitting Time or Obesity: Which Came First? Bidirectional Association in A Longitudinal Study of 31,787 Australian Adults. *Obesity* 22, 2126–2130.DOI: https://doi.org/10.1002/oby.20817

Pulsford, R.M., Stamatakis, E., Britton A.R., Brunner, E.J., and Hillsdon, M.M. 2013. Sitting Behavior and Obesity: Evidence from the Whitehall II Study. *The American Journal of Preventive Medicine* 44: 132–138. DOI: https://doi.org/10.1016/j.amepre.2012.10.009

Radford, J. and Joseph, K. 2020. Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. *Frontiers in Big Data* 3, 18–18. DOI: https://doi.org/10.3389/fdata.2020.00018

Reile, R., Baburin, A., Veideman, T., and Leinsalu, M. 2020. Long-Term Trends in the Body Mass Index and Obesity Risk in Estonia: An Age-Period-Cohort Approach. *The International Journal of Public Health* 65(6), 859−869. DOI: https://doi.org/10.1007/s00038-020-01447-7

Reile, R. and Leinsalu, M. 2019. Factors Associated with Improving Diet and Physical Activity Among Persons with Excess Body Weight. *The European Journal of Public Health* 29(6), 1166–1171. DOI: https://doi.org/10.1093/eurpub/ckz170

Reile, R., Tekkel, M. and Veideman, T. 2019. *Eesti täiskasvanud rahvastiku tervisekäitumise uuring 2018*. Tallinn: Tervise Arengu Instituut.

Reile, R. and Veideman, T. 2021. *Eesti täiskasvanud rahvastiku tervisekäitumise uuring 2020*. Tallinn: Tervise Arengu Instituut.

Rios-Julian, N., Alarcon-Paredes, A., Alonso, G.A., et al. 2017. Feasibility of a Screening Tool for Obesity Diagnosis in Mexican Children from a Vulnerable Community of Me'Phaa Ethnicity in the State of Guerrero, Mexico. *2017 Global Medical Engineering Physics Exchanges/Pan American*. IEEE: Tuxtla-Gutierrez, Mexico, 1–6. DOI: https://doi.org/10.1109/GMEPE-PAHCE.2017.7972105

Selya, A.S. and Anshutz, D. 2018. Machine Learning for the Classification of Obesity from Dietary and Physical Activity Patterns. In Giabbanelli, P.J., Mago, V.K., and Papageorgiou, E.I. (eds.) *Advanced Data Analytics in Health*. Cham: Springer, 77–97. DOI: https://doi.org/10.1007/978-3-319-77911-9_5

Seo, D.C. and Li, K. 2010. Leisure-Time Physical Activity Dose-Response Effects on Obesity Among US Adults: Results from the 1999-2006 National Health and Nutrition Examination Survey. *The Journal of Epidemiology and Community Health* 64, 426–431.DOI: http://dx.doi.org/10.1136/jech.2009.089680

Shmueli, G. 2010. To Explain or to Predict? *Statistical Science* 25(3), 289–310. DOI: https://doi.org/10.1214/10-STS330

Stamatakis, E., Hirani, V., and Rennie, K. 2009. Moderate-to-vigorous Physical Activity and Sedentary Behaviours in Relation to Body Mass Index-Defined and Waist

Circumference-Defined Obesity. *The British Journal of Nutrition* 101(5), 765–773. DOI: https://doi.org/10.1017/S0007114508035939

Tekkel, M. and Veideman, T. 2011. *Eesti täiskasvanud rahvastiku tervisekäitumise uuring, 2010*. Tallinn: Tervise Arengu Instituut.

Tekkel, M. and Veideman, T. 2013. *Eesti täiskasvanud rahvastiku tervisekäitumise uuring, 2012*. Tallinn: Tervise Arengu Instituut.

Tekkel, M. and Veideman, T. 2015. *Eesti täiskasvanud rahvastiku tervisekäitumise uuring, 2014*. Tallinn: Tervise Arengu Instituut.

Tekkel, M. and Veideman, T. 2017. *Eesti täiskasvanud rahvastiku tervisekäitumise uuring, 2016*. Tallinn: Tervise Arengu Instituut.

Tekkel, M., Veideman, T., and Rahu, M. 2010. Changes over Fourteen Years in Adult Obesity in Estonia: Socioeconomic Status and Use of Outpatient Health Services. *Central European Journal of Public Health* 18(4), 186–191. DOI: https://doi.org/10.21101/cejph.a3588

Thamrin, S.A., Sidik, A.D, Hedi, K., Armin, L., and Sudirman; N. 2021. Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018. *Frontiers in Nutrition* 8, 669155. DOI: https://doi.org/10.3389/fnut.2021.669155

Townshend, T. and Lake, A. 2017. Obesogenic Environments: Current Evidence of the Built and Food Environments. *Perspectives in Public Health* 137(1), 38–44. DOI: https://doi.org/10.1177/1757913916679860

Vaishya, R., Javaid, M., Khan, I.H., and Haleem, A. 2020. Artificial Intelligence (AI) Applications for COVID-19 Pandemic. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* 14(4), 337–339. DOI: https://doi.org/10.1016/j.dsx.2020.04.012

Wanner, M., Martin, B., Autenrieth, C. et al. 2016. Associations Between Domains of Physical Activity, Sitting Time, and Different Measures of Overweight and Obesity. *Preventive Medicine Reports* 3, 177–184. DOI: https://doi.org/10.1016/j.pmedr.2016.01.007

Wanner, M., Richard, A., Martin, B., Faeh, D., and Rohrmann, S. 2017. Associations Between Self-Reported and Objectively Measured Physical Activity, Sedentary Behavior and Overweight/Obesity in NHANES 2003–2006. *The International Journal of Obesity* 41, 186–193. DOI: https://doi.org/10.1038/ijo.2016.168

Wareham, N.J., van Sluijs, E.M., and Ekelund, U. 2005. Physical Activity and Obesity Prevention: A Review of the Current Evidence. *Proceedings of the Nutrition Society* 64(2), 229–247. DOI: https://doi.org/ 10.1079/PNS2005423

Webber, L et al. 2012. Modelling Obesity Trends and Related Diseases in Eastern Europe. *Obesity Reviews* 13(8), 744–751. DOI: https://doi.org/10.1111/j.1467-789X.2012.00999.x

Willetts, M., Hollowell, S., Aslett, L. et al. 2018. Statistical Machine Learning of Sleep and Physical Activity Phenotypes From Sensor Data in 96,220 UK Biobank Participants. *Scientific Reports* 8, 7961. DOI: https://doi.org/10.1038/s41598-018-26174-1

World Health Organization 2021. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. (accessed on 27 September 2021)

World Health Organization 2000. *Report of a WHO Consultation: Obesity: Preventing and Managing the Global Epidemic*. Geneva: World Health Organization.

Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., and Keane, J. 2009. Comparing Data Mining Methods with Logistic Regression in Childhood Obesity Prediction. *Information Systems Frontiers* 11:449–60. DOI: https://doi.org/10.1007/s10796-009-9157-0

Zheng, Z. and Ruggiero, K. 2017. Using Machine Learning to Predict Obesity in High School Students. *Proceedings of the 2017 IEEE International Conference on Bioinformatics*

and Biomedicine (BIBM), 2132–2138. DOI: https://doi.org/10.1109/BIBM.2017.8217988

Zhou, M., Fukuoka, Y., Goldberg, K. et al. 2019. Applying Machine Learning to Predict Future Adherence to Physical Activity Programs. *BMC Medical Informatics and Decision Making* 19, 169. DOI: https://doi.org/10.1186/s12911-019-0890-0

8 Appendices

8.1 Glossary

Accuracy – the proportion of all cases predicted correctly by the model from the total number of predictions

Area under receiver operating characteristic curve (AUC, also known as C-statistic, or simply "the area under the curve") – a metric that measures the ability of a classifier to distinguish between different classes; it is used as a summary metric of the ROC curve

AUC – see "Area under receiver operating characteristic curve"

BMI – see "Body mass index"

Body mass index – a common metric that is used to group individuals into weight categories; it is calculated as the person's body weight in kilograms divided by the person's squared standing height in meters and measured in the units of kg/m^2

CI – see "Confidence interval"

Confidence interval – a range of likely values of the studied parameter, at a certain confidence level (most commonly 95% as also in this study)

Decision tree – a common classification algorithm that assigns cases to specific classes by determining the probability thresholds derived from input data (DeGregori et al 2017, 672)

DT – see "Decision tree"

Explanatory modelling – the construction and use of statistical models with an aim to test causal explanations (Shmueli 2010, 290)

F1-score (also known as F-score) - the harmonic mean of the model's recall and precision

False negative (FN) – a result that wrongly indicates the absence of certain condition or attribute

False positive (FP) – a result that wrongly indicates the presence of certain condition or attribute

FN – see "False negative"

FP – see "False positive"

HBEAPS – see "Health Behaviour Among Estonian Adult Population survey"

Health Behaviour Among Estonian Adult Population survey – a nationwide population-based cross-sectional survey, conducted biannually since 1990 by the Estonian National Institute for Health Development

Hyperparameter – a model parameter the value of which is set before the model training process beings

k-nearest neighbors (KNN, also k-NN) – a common classification algorithm that assigns a class label to a new instance based on the class assignment of its k nearest neighbors with predictor variables least distant from those of the new instance (Colmenarejo 2020, 8)

KNN – see "k-nearest neighbours"

Logistic regression – a common linear classification model to predict binary outcomes (in case of binary logistic regression) or multi-class outcomes (in case of multivariate logistic regression) (Thamrin et al 2021, 3)

LR – see "Logistic regression"

Naïve Bayes – a classification algorithm belonging to a bigger family of Bayesian classifiers that predict class membership by probabilities (Zhang 2009, 452)

NB – see "Naïve Bayes"

Obesity – according to the single-cut definitions of the BMI categories by the World Health Organization (2000), a weight category with BMI $\geq$ 30.0 kg/m^2. Obesity is sometimes (although not in this thesis) further divided into three sub-categories: obesity class I (BMI = 30.0–34.9 kg/m^2), obesity class II (BMI = 35.0–39.9 kg/m^2), and obesity class III (BMI $\geq$ 40.0 kg/m^2) (Chatterjee et al 2020; Jindal et al 2018)

Odds ratio (OR) – the ratio of the odds of the presence of certain condition or attribute in one group to the odds of that in the other

OR – see "Odds ratio"

Overweight – according to the single-cut definitions of the BMI categories by the World Health Organization (2000), a weight category with BMI = 25.0–29.9 kg/m^2. Importantly, throughout this thesis the category "overweight," unless stipulated otherwise, refers to "BMI $\geq$ 25.0 kg/m^2", i.e. it comprises both overweight and obese categories (BMI $\geq$ 30.0 kg/m^2) as they are defined by the World Health Organization

Precision – the proportion of true positives among all instances predicted as positive by the classifier

Predictive modelling – the construction and use of models with an aim to produce predictions (Shmueli 2010, 291)

Random forest – an ensemble method that is based on the construction of many decision trees

Recall (also sensitivity or the true positive rate) – the proportion of true positives that are correctly identified by the classifier from all positive cases

Receiver operating characteristic curve (ROC curve) – a probability curve that plots the true positive rate against false positive rate at various threshold values

RF – see “Random forest”

ROC curve – see “Receiver operating characteristic curve”

Sensitivity – see “Recall”

Support vector machine – a machine learning method that is based on the construction of a so-called “hyperplane” with maximal margin from the predictor variables (Colmenarejo 2020, 9)

SVM – see “Support vector machine”

TN – see “True negative”

TP – see “True positive”

True negative (TN) – a result that correctly indicates the absence of certain condition or attribute

True positive (TP) – a result that correctly indicates the presence of certain condition or attribute

8.2 License

Non-exclusive licence to reproduce thesis and make thesis public

I, **Toomas Gross**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Explanatory and Predictive Modelling in the Study of Overweight and Obesity: The Example of Health Behaviour Among Estonian Adult Population

supervised by Rajesh Sharma

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Toomas Gross

20/12/2021