Tampere University

Heidi Mikkola

PREDICTION OF CHILDREN’S OVERWEIGHT USING SUPERVISED LEARNING

Master’s Thesis
Faculty of Information Technology and Communication Sciences
April 2023

ABSTRACT

Heidi Mikkola: Prediction of children's overweight using supervised learning
Master's Thesis
Tampere University
Master's Degree Programme in Computational Big Data Analytics
April 2023

Being overweight is a massive problem all over the world, and it affects developed countries also, including Finland. The aim of this thesis is to investigate whether is it possible to predict if a child is at risk to become overweight later in life using longitudinal data on a child's BMI. The data that was used in this thesis were collected in the Pirkanmaa area in Finland and it contains anthropometric measurements from 4223 children that were born in 1974, 1981, 1991, and 2001. The measurements were taken from birth up to 15 years of age, except for children who were born in 2001. They were measured only up to 11 years of age. Supervised learning, specifically discriminant analysis was used to cluster children into two groups depending are they at a risk to become overweight at 15 years of age. Being overweight was defined as BMI equal to or over 25. Randomly selected 70% of the whole data was used as a training set and the remaining 30% of the data was used as a test set. The mixed models were created by using the training set and it was applied to the test set which contains data from 1 year of age up to 7 years of age. This thesis used and compared three different prediction approaches, marginal, conditional, and random effect predictions, to predict children who are at risk to become overweight at 15 years of age. The results were compared to actual values. Marginal and conditional predictions gave similar results but the random effect prediction approach seemed to work worse than marginal and conditional predictions. The results indicated that the older a child is the easier is to see if the child will be overweight. Thus, it is really difficult to predict possible future overweight with very young children whose maximum age is 2 years. In the girls' and boys' groups were no big differences, although there could be seen as slightly better predictable in the girls' group with very young children. Nevertheless, the improvement of predictability was better in the boys' group than in the girls' group.

Keywords: supervised learning, discriminant analysis, clustering, body mass index, longitudinal data

TIIVISTELMÄ

Heidi Mikkola: Lasten ylipainon ennustaminen käyttäen ohjattua oppimista
Pro Gradu -tutkielma
Tampereen yliopisto
Master's Degree Programme in Computational Big Data Analytics
Huhtikuu 2023

Ylipaino on valtava ongelma ympäri maailmaa, ja se koskettaa myös kehittyneitä valtioita, kuten Suomea. Tämän työn tarkoituksena on tutkia, pystyykö lasten BMI-tiedoista ennustamaan, tulevatko he olemaan ylipainoriskissä myöhemmin elämässään. Aineistona käytettiin Suomessa Pirkanmaalla kerättyä pitkittäisaineistoa, joka sisälsi 4223 lapsen antropometrisia tietoja. Aineistossa oli kerätty vuosina 1974, 1981, 1991, 1995 ja 2001 syntyneiltä lapsilta muun muassa pituutta ja painoa. Dataa oli kerätty syntymästä 15-vuotiaaksi asti, lukuunottamatta 2001 syntyneitä, joilla mittauksia oli tehty vain 11-vuotiaaksi asti. Tässä työssä käytettiin ohjattua oppimista ja diskriminanttianalyysiä klusterointiin. Lapset luokiteltiin kahteen eri ryhmään, ylipainoisiin ja normaalipainoisiin, sen perusteella tulevatko he olemaan 15-vuotiaina ylipainoisia vai ei. Ylipaino määriteltiin painoindeksin ollessa 25 tai sen yli. Opetusjoukkona käytettiin satunnaisesti valittua 70 prosenttia koko aineistosta, ja testijoukkona loppua 30 prosenttia. Opetusjoukon avulla mallinnettuja sekamalleja käytettiin testijoukkoon, jossa oli lapsilta dataa 1-vuotiaasta 7-vuotiaaksi saakka. Testijoukon lapset luokiteltiin kahteen ryhmään heidän todennäköisyyden olla 15-vuotiaina ylipainoisia perusteella. Tässä työssä käytettiin ja verrattiin kolmea eri ennustemenetelmää, marginaaliennuste, ehdollinen ennuste sekä satunnaisvaikutusennuste. Tuloksia verrattiin oikeisiin arvoihin ja tulokset osoittivat, että marginaaliennuste sekä ehdollinen ennuste antoivat samanlaisia tuloksia, mutta satunnaisvaikutusennuste toimi näitä kahta muuta ennustetapaa huonommin. Tuloksista nähtiin, että mitä vanhempi lapsi on, sitä helpompi on nähdä tuleeko hänestä ylipainoinen vai ei. Aivan pieniltä lapsilta, joiden ikä oli maksimissaan 2 vuotta, oli vaikea ennustaa mahdollista tulevaa ylipainoa. Tyttöjen ja poikien välillä ei ollut suuria eroja, joskin tytöillä oli aivan nuorena havaittavissa hieman parempaa ennustettavuutta kuin pojilla, mutta poikien ennustettavuus parani selkeämmin ajan myötä.

Avainsanat: ohjattu oppiminen, diskriminanttianalyysi, klusterointi, painoindeksi, pitkittäisaineisto

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Being overweight is a major health problem all over the world. It is a serious disease, and it affects negatively the quality and duration of life. Being overweight and obese does not affect only adults but also children and adolescents. Childhood obesity has increased greatly over the past three decades. World Health Organization (WHO) has reported that 42 million children under the age of 5 were overweight or obese in 2016 [1]. More than 190 million children and adolescents aged 5 to 19 were reportedly obese in 2019 [2]. It is a big health problem that the growth of the amount of overweight children keeps growing all the time and also during the COVID-19 pandemic childhood obesity has increased [3]. That is very alarming, and it has a significant effect on the whole of society. In Finland overweight is also a problem. World Obesity Federation (WOF) has collected data about childhood obesity in all countries of the world and in 2016 15.6% of Finnish boys aged 5-9 were obese and the proportion of Finnish girls was 7.8% [2]. The proportions are a little bit lower for children aged 10-19: 10.8% of boys and 4.5% of girls were obese in Finland [2]. Worldwide 20.6% of children aged 5-9 were overweight in 2019 which means 131 million children [4].

There are a lot of negative health consequences of being overweight in childhood. It has a massive impact on physical health. Children who have overweight are at increased risk for many serious diseases. Cardiovascular disease, type 2 diabetes, asthma, high cholesterol, skin conditions, menstrual abnormalities, and orthopedic problems are examples that people with overweight have more often than people with normal weight [5].

Being overweight can also cause many problems to psychological health. It has been studied that childhood obesity has a negative impact on children's social and emotional well-being and self-esteem also [5]. Being overweight can cause social problems, which can show up for example that children with overweight are more often bullied and tend to have fewer friends than children with normal weight. Being overweight also has an impact on education because overweight children have more likely problems with school. Some of these problems, both physical and psychical problems, can continue throughout adulthood and probably throughout the whole life [5].

Overweight is unquestionably a massive problem for society and it has many negative consequences. Therefore, it would be necessary to try to decrease childhood obesity and identify the causes that lead to being overweight. The aim of this thesis is to investigate how well can we predict those children who are later overweight based on body mass index before age of 7 using supervised learning.

This thesis uses supervised learning to recognize these children who have a risk to be overweight later in life. Supervised learning is one approach to machine learning, which is a subfield of artificial intelligence [6]. Artificial intelligence means a machine's capability to mimic human intelligence, like predicting, inferring, learning, and creating [7]. It is an important part of society's digital transformation and that is why it is commonly used and the target of many investigations. Artificial intelligence is all along more and more part of our everyday life. Machine learning enables machines to learn from data without specific programming [6]. It uses a massive amount of data to learn a model, which generates results and predictions. If a machine gets input-output pairs and learns based on these pairs, it is called supervised learning [6] and it is a method that is used in this thesis.

Supervised learning is commonly used because it is fast and cheap. It is meaningful to investigate how well is it possible to predict children's BMI later in life using supervised learning because it could help to prevent overweight. This thesis applies supervised learning to predict children which are overweight later in life. Consequently, this thesis wants to classify children into two different groups: overweight and not overweight. This thesis uses training data to learn a model which is then applied to the data in the test set and predicts whether those children are going to be overweight or not. Thus, the thesis uses classification techniques to predict those children with a risk of being overweight.

We use quite an extensive longitudinal data set in this thesis. It contains anthropometric measurements of Finnish children who are born between 1974 and 2001. Children's height and weight are measured when children are born and then afterward in seven routine health check-up times, up to 15 years of age. Data includes measures about height and weight and this thesis uses these measurements to calculate BMI and predict overweight. We use randomly selected two-thirds of the data as a training set and the remaining third of the data as a test set.

This longitudinal children's BMI data set is already investigated in earlier studies. Nummi et al. [8] are focused on the trajectory analysis and the aim was to identify and model BMI development. This thesis focuses on discriminant analysis and classifying children who are at risk to become overweight. Overweight and obesity are common topics to investigate, but as long as overweight is a massive worldwide health problem and it is possible to get some new methods to prevent overweight by investigating, it is useful to do research on that topic. It would be very beneficial if we could find a good statistical method to classify before 7 years of age those children who are at risk of becoming overweight or obese. In that way, it is possible to focus on these people and try to prevent overweight or obesity.

Chapter 2 introduces how can be defined children's overweight and obesity. Chapter 3 describes which methods are applied in this thesis. Describing data and software which was utilized in this thesis are reported in chapter 4. After that, chapter 5 introduces the results, and finally, chapter 6 is discussed in the conclusions of this thesis and considered what could be researched in the future about this subject.

2. LITERATURE REVIEW

Obesity is defined as abnormal or excessive fat accumulation in adipose tissue, which may impair health. Obesity is defined as a condition where there is too much adipose tissue in a body. Adipose tissue is storing fat and it is the human's body energy storage. If there is too much fat in adipose tissue it means overweight. Getting more energy from food than it is expended is the general reason to weight gain.

The most used measure for obesity is body mass index (BMI). It is a simple index to estimate body fat using a relationship between weight and height. Exactly definition for BMI, therefore, is the weight in kilograms divided by the square of the height in meters:

$$\text{BMI} = \frac{\text{Weight}}{\text{Height}^2}. \tag{2.1}$$

BMI works well, especially for adults and WHO has defined BMI categories that can be seen in table 2.1. When BMI is 18.5-24.9 it means healthy weight and under that it means underweight. When BMI is 25.0-29.9 it signifies overweight. Obesity is defined as when BMI increases over 30.0. Obesity class II denotes severe obese and obesity class III means very severely obese.

BMI has applied also for children and adolescents and it is calculated with the same formula as an adult's BMI. There is a different interpretation of children's BMI because children are still growing and BMI categories are not so straightforward. Speed of growth varies along with age and gender, and girls and boys have different amounts of fat in their

Table 2.1. *BMI categories*

BMI	Nutritional status
Below 18.5	Underweight
18.5-24.9	Normal weight
25.0-29.9	Overweight
30.0-34.9	Obesity class I
35.0-39.9	Obesity class II
Above 40	Obesity class III

bodies, therefore we must consider gender and model girls and boys separately to get better results. The world health organization (WHO) has defined BMI-for-age as boys and girls. It means that children's and adolescents' BMI is calculated with the same formula as BMI for adults, but the BMI is then compared with percentiles to get to know which nutritional status category a child is counted as. Overweight, in children the age of 5-19 years, is defined BMI-for-age as greater than 1 standard deviation above the median. If it is greater than 2 standard deviations, it signifies obesity. With younger children, under the age of 5, there is used weight-for-height and if that is greater than 2 standard deviations above the median, it means overweight, and greater than 3 standard deviations above the median identify obesity.

BMI is not the only indicator of overweight and obesity. Other indicators are for example waistline measurements, waist-hip ratio, waist-height ratio, and total body fat [9]. Nevertheless, BMI is indisputably the most used indicator of overweight and obesity, even though it is not entirely flawless. There are investigations about BMI's capability and how good an indicator it is, and also criticism has been given. Karchynskaya et. al [10] have investigated how good an indicator BMI is. They have shown that BMI and the proportion of fat differ; one reason is that BMI does not consider muscles and bones. It takes only into account weight, and not where the weight is come from. This can lead to misclassification. Despite these issues, BMI is used in this thesis because it is the simplest and maybe the most used indicator for overweight and obesity. Gender is also a significant factor when considering BMI, and Rothman [11] has presented that BMI and percentage of body fat are different for men and women. Thus, also in this thesis boys and girls are examined separately.

3. THEORETHICAL BACKGROUND

3.1 Supervised learning

Supervised learning is an approach to machine learning. Machine learning, in turn, is a subfield of artificial intelligence, and the basic idea of machine learning is to enable machines to learn using previous data or experiences without any specific programming [6]. Machine learning needs a lot of data, named training set, to learn a model which then gives predictions and results to data in a test set based on training data [12]. The training set consists of dependent variables and target variables. When a computer algorithm is trained to classify new data using labeled data, it is called supervised learning. It uses input-output pairs to generate an algorithm and learn a mapping function between these input and output variables [6]. This generated algorithm is then used to classify data and to give labels to predictive variables in the new data. If the (x_i, y_i) are the data pair, where x_i is the input variable and y_i is the output variable, and $f(x)$ is the mapping function, it can be shown that the learned model is

$$y_i = f(x_i) + \epsilon_i, \tag{3.1}$$

where the ϵ_i is a random error.

Supervised learning can be separated into two approaches: classification and regression [13]. Classification uses training data to learn an algorithm or model, which is then used to classify test data into specific categories. Regression is used to understand the relationship between dependent and independent variables and find correlations between them. The other difference between classification and regression is that classification is applied for discrete values and regression is applied for continuous values [13].

3.2 Classification

Classification is very commonly used in many fields of science and scientific studies. Classification is one type of supervised learning, and it is the process of labeling data based on the properties of the data. The main purpose of classification is to get figure out which category a new object belongs to [13]. To get to know which category unclassified objects belong to the model must be created, and it gives results for classification. The model can be constituted using training data.

Thus, classification predictive modeling is an approach to approximate mapping function based on input variables to get discrete output variables. Classification can be separated roughly into two main categories: discriminant analysis and cluster analysis [14]. In a situation, where data and groups of that data are known a priori and then observations of measurements whose group membership is not known are wanted to determine, discriminant analysis is used. Discriminant analysis can be named also learning with a teacher [14]. In cluster analysis groups are not known a priori, but the purpose is to define groups based on the data. Thus, entities inside one group have more similar features than other entities in other groups. Cluster analysis can be called also learning without a teacher [14]. This thesis focuses on discriminant analysis.

3.3 Discriminant analysis

3.3.1 Main idea of discriminant analysis

Discriminant analysis is one technique to classify entities. The main idea is to classify data in which group membership is unknown, called also test data, into groups based on data, called training data, in which group membership is known. One entity of test data belongs to one of G possible classes. Therefore, the basic problem is to classify random vector $\mathbf{y}$ to one of the k populations. Gnanadesikan et al. [14] explained that discriminant analysis can be separated into two stages. First, G groups have to separate clearly by using training data, and it is the first stage. In the second stage, test data has to classify into these groups.

3.3.2 Longitudinal discriminant analysis

If there is a dataset that contains repeated measurements of one entity over some time and some future point is wanted to know, that kind of problem can be solved using longitudinal discriminant analysis [15]. In case of longitudinal data and when $\mathbf{Y}$ denotes a random vector, let $\mathbf{Y} = (Y_{r,1}, ..., Y_{r,n})^\top$ be a vector of longitudinal data of the r marker (variable). Rth marker has observed at time points $\mathbf{t} = (t_1, t_2, ..., t_n)^\top$.

Let $\rho_{g,i}(t)$ be the probability that the entity i belongs to the specific group g, from all the groups G, at time t. The aim is to estimate $\rho_{g,i}(t)$ and use it to classify entities into groups. The statistical model must be used to approximate this probability. It is assumed that training data includes entities that belong to all groups G in the future, and it is known. Thus, it is possible to learn a model which categorizes entities into all groups G.

Data can include different predictor variables as known as markers, and sometimes it is necessary to use many markers in longitudinal discriminant analysis methodologies to get proper results. These markers may be continuous or discrete variables and hence some methods of longitudinal discriminant analysis utilize a mixed model methodology. Komarek et al. [16] have introduced a multivariate linear mixed model (MLMM) with a normal mixture in the random effects distribution.

3.4 A multivariate linear mixed model with a normal mixture in the random effects distribution

Mixed models are useful in a longitudinal study where measurements are collected repeatedly on the same unit. If a discriminant analysis is based on mixed models, it allows every unit not to have to measure every time [15] and that is a benefit of mixed models. Linear mixed models are useful to handle missing data where the data are missing at random [17]. In longitudinal studies, it is common that every unit is not possible to measure at every time point and therefore mixed models are beneficial in longitudinal studies. The mixed model is a statistical model, which contains both fixed effects and random effects. Fixed effects are used to model an average behavior with all units and random effects model a variation around the specific parameter. In the other words, fixed effects refer to the population average, and random effects model the variance of the particular parameters at different times and therefore it is subject specific. Random effects are parameters that are themselves random variables whereas fixed effects are not random [18].

Komarek et al. [16] have defined a multivariate linear mixed model (MLMM) for the rth marker as

$$\mathbf{Y}_{i,r} = \mathbf{X}_{i,r}\boldsymbol{\alpha}_r + \mathbf{Z}_{i,r}\mathbf{b}_{i,r} + \boldsymbol{\epsilon}_{i,r}, \tag{3.2}$$

where $(i = 1, ..., N, r = 1, ..., R)$ and $\mathbf{Y}_{i,r}$ is a response vector for marker r on the ith unit and $\boldsymbol{\alpha}_r$ is a vector of fixed effects for marker r and $\mathbf{X}_{i,r}$ is a covariate matrix for the fixed effects in a model for marker r. A vector of individual random effects for marker r is $\mathbf{b}_{i,r}$ and $\mathbf{Z}_{i,r}$ is a covariate matrix for the random effects in a model for marker r. A vector of random errors for marker r on the ith unit is $\boldsymbol{\epsilon}_{i,r}$ and errors are assumed to be independent and normally distributed, $\boldsymbol{\epsilon}_{i,r} \sim N(\mathbf{0}, \sigma_r^2\mathbf{I})$ where σ_r^2 stands as the residual variance of the rth marker.

The random effect vector $\mathbf{b}_{i,r} = (b_{i,1}, ..., b_{i,R})$ can be used to define the possible correlation between repeated measurements of either the same marker or between markers. A vector of ith subject-specific random effects to all markers R is denoted by $\mathbf{b}_i$. In many studies, random vector $\mathbf{b}_i$ is assumed to follow a normal distribution. Hughes et al. [15] have explained that it can lead to the affected performance of the discriminant procedure. Assuming the random vector follows a normal distribution can cause the model misspecification of the random effects distributions and therefore model estimates can thus be biased [15]. To tackle misspecification, a normal mixture in the random effect distribution is used here. Hence, it is assumed that

$$\mathbf{b}_i \sim \sum_{k=1}^{K} w_k N(\boldsymbol{\mu}_k, \mathbf{D}_k), \tag{3.3}$$

where N means multivariate normal distribution with the mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{D}_k$. Unknown model parameters are $\boldsymbol{\mu}_k$, $\mathbf{D}_k$, and w_k which denotes non-negative mixture weights. The number of mixture components K is assumed to be pre-specified. The mixture distribution is compounded from individual distributions which are called mixture components. Each mixture component has a weight or probability which is related to the component and that is called mixture weight.

The mixture means, covariance matrices and mixture weights are the model parameters that must be estimated for each prognostic group. The vector of these unknown parameters can be denoted as

$$\boldsymbol{\theta}^g := (\mathbf{w}^g, \boldsymbol{\mu}_1^g, ..., \boldsymbol{\mu}_K^g, \mathbf{D}_1^g, ..., \mathbf{D}_K^g). \tag{3.4}$$

Also, unknown fixed effect parameters must be estimated for each group. These can be written as

$$\boldsymbol{\psi}^g := (\boldsymbol{\alpha}_1^g, ..., \boldsymbol{\alpha}_R^g, \sigma_1^g, ..., \sigma_R^g), \tag{3.5}$$

where $\boldsymbol{\alpha}_r^g = (\boldsymbol{\alpha}_1^g, ..., \boldsymbol{\alpha}_R^g)$ is a vector of fixed effects, and the vector of residual variances is $\boldsymbol{\sigma}_r^g = (\sigma_1^g, ..., \sigma_R^g)$. To estimate these unknown parameters, this thesis uses Bayes' theorem which is introduced in the next subsection.

3.4.1 Bayes' theorem

In discrimination analysis the main purpose is to use model parameters, mixture-related parameters $\boldsymbol{\theta}^g$, and fixed effects regression coefficients $\boldsymbol{\psi}^g$ which are estimated for all groups, to classify objects using their historical data which is known. Bayes' theorem is utilized in this kind of case.

Bayes' theorem is applied in Bayesian statistical methods. Bayes' theorem is used to calculate the posterior probability of an event based on prior probability and the observed data [19]. In the other words, it is used for calculating conditional probability which means a probability that an event is true only if another event is true. In practice, the posterior probability is calculated using prior probability and the likelihood [15].

Allocation of the new subject into one of the groups G is represented by a random vector $\mathbf{U}_i$. This thesis assumes that $\mathbf{U}_i$ contains values only up to time t < T and value $u_i(t)$ of $\mathbf{U}_i$ is not observed and the purpose is to predict it estimating $\mathcal{P}_{i,g} = \mathcal{P}_{i,g}(t)$, which is a probability that $u_i(t)$ belongs to the group. Let $\pi = P(\mathbf{U}_i = g)$ be a prior probability to $\mathbf{U}_i$ belongs the group, and it can be called also a prior group probability. Using prior group probability, we get the posterior probability for each group denoting that u_i belongs to the group. That takes a form

$$\mathcal{P}_{i,g}(\boldsymbol{\psi}, \boldsymbol{\theta}) = P(u_i = g|\boldsymbol{\psi}, \boldsymbol{\theta}, \mathbf{y_i}) = \frac{\pi_g f_{i,g}}{\sum_{l=1}^{G} \pi_l f_{i,l}}, \tag{3.6}$$

where $(i = 1, ..., N, g = 1, ..., G)$ and π is a prior probability and $f_{i,g}$ is a prior distribution. Thus, the estimated group probability is a combination of the prior probability and the prior distributions.

The Bayesian approach can be done using Markov Chain Monte Carlo (MCMC) simulation because it is a technique that is developed for getting a solution of the posterior probability.

3.4.2 Markov Chain Monte Carlo

Monte Carlo is a technique to estimate possible results of uncertain events. It is a method that randomly samples again and again random values using probability distribution [20]. The Monte Carlo technique approximates the desired quantity and it is used in cases where estimated quantities are impossible or difficult to compute exactly.

Markov Chain in turn is a process for generating a sequence of random variables whose probabilities depend on the value of the recent prior variable and only on that [21]. Let $(X_1, ..., X_n)$ be a sequence of random variables. Calculating the probability that a future variable X_{n+1} is x depends only on a random variable X_n. Thus, earlier variables do not affect the state of a random variable X_{n+1} [22]. This can be denoted as follows

$$P(X_{n+1} = x|X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = x|X_n = x_n). \tag{3.7}$$

Markov Chain Monte Carlo (MCMC) is a combination of these two methods [21]. Consequently, MCMC allows random sampling of high-dimensional probability distributions.

3.5 Marginal, conditional and random effects prediction

Marginal, conditional, and random effects predictions are different approaches to getting the final group probabilities and they are used in longitudinal discriminant analysis based on mixed models [23]. The main idea of these predictions is to compare group probabilities, which are based on historical data, and a longitudinal profile of the new subject and then set the subject into the group in which $\mathcal{P}_{i,g}$ is the highest. Accordingly, the new subject is classified into a group whose longitudinal profile is closest to the subject's profile. Morrel et al. [23] have shown that marginal prediction is based on the marginal distribution of observed markers, conditional prediction is based on the conditional distribution of observed markers given suitable predictors of random effect, and random effects prediction is based on the distribution of random effects.

Marginal prediction is commonly applied in studies that use longitudinal data and is based on the marginal distribution of observed markers. In this prediction, the random effects are integrated out because it is focusing on the mean evolution of the marker over time [15]. In the other words, model parameters depend only on the values of the observable longitudinal markers of the new subject. In marginal prediction, for observed values $\mathbf{y}_1 = (y_{1,1}, ..., y_{1,n}), ..., \mathbf{y}_R = (y_{R,1}, ..., y_{R,n})$ of the longitudinal markers $\mathbf{Y} = (Y_{r,1}, ..., Y_{r,n})^\top$ for a subject from the prognostic group g, posterior probabilities can be calculated as

$$\mathcal{P}_{i,g}^{marg}(\boldsymbol{\psi}, \boldsymbol{\theta}) = \frac{\pi_g f_g^{marg}(\mathbf{y}_{i,1}, ..., \mathbf{y}_{i,R}; \boldsymbol{\psi}^g, \boldsymbol{\theta}^g)}{\sum_{\overline{g}=0}^{G-1} \pi_{\overline{g}} f_{\overline{g}}^{marg}(\mathbf{y}_{i,1}, ..., \mathbf{y}_{i,R}; \boldsymbol{\psi}^{\overline{g}}, \boldsymbol{\theta}^{\overline{g}})}, \tag{3.8}$$

where $\boldsymbol{\psi}^g$ and $\boldsymbol{\theta}^g$ are the unknown parameters introduced in 3.4 and 3.5, and f_g^{marg} denotes marginal density which is defined using densities of conditional and random effect. Accordingly,

$$f_g^{marg}(\mathbf{y}_1, ..., \mathbf{y}_R; \boldsymbol{\psi}^g, \boldsymbol{\theta}^g) = \int f_g^{cond}(\mathbf{y}_1, ..., \mathbf{y}_R | \mathbf{b}; \boldsymbol{\psi}^g) f_g^{ranef}(\mathbf{b}; \boldsymbol{\theta}^g) \mathrm{d}b. \tag{3.9}$$

Conditional prediction is based on the conditional distribution of observed markers. It is considered a suitable predictor of individual random effects. Conditional prediction focuses on the subject-specific evolution of the markers over time given the estimated values of individual random effects. Conditional prediction does not take into account variability in the estimation of the individual random effects [16]. Hence, a conditional density can be denoted as

$$f_g^{cond}(\mathbf{y}_1, ..., \mathbf{y}_R | \mathbf{b}; \boldsymbol{\psi}^g) = \prod_{r=1}^{R} \prod_{j=1}^{n_r} p_r(\mathbf{y}_{r,j} | \mathbf{b}; \boldsymbol{\psi}^g), \tag{3.10}$$

where $p_r(\mathbf{y}_{r,j} | \mathbf{b}; \boldsymbol{\psi}^g)$ is a density of the exponential family distribution assumed for the rth marker, $\mathbf{b}$ is a vector of random effects and $\boldsymbol{\psi}^g$ is a parameter vector which was introduced in 3.5.

Random effects prediction is founded on the distributions of the individual random effects [16]. It focuses on the subject-specific evolution of longitudinal profiles such as conditional prediction. Komarek et al. [16] have introduced that random effect prediction uses only the latent characteristics of the subjects with the between-subjects variability. A random effects density for group g is

$$f_g^{ranef}(\mathbf{b};\boldsymbol{\theta}^g) = \sum_{k=1}^{K^g} w_k^g p_k(\mathbf{b};\boldsymbol{\mu}_k^g,\mathbf{D}_k^g), \tag{3.11}$$

where $p_k(\mathbf{b};\boldsymbol{\mu}_k^g,\mathbf{D}_k^g)$ is a density of multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and a covariance matrix $\mathbf{D}_k$.

Using these densities we get group probabilities for the random effects and conditional predictions

$$\mathcal{P}_{g,i}^{ranef}(\mathbf{b}_i^0,...,\mathbf{b}_i^{G-1},\boldsymbol{\psi},\boldsymbol{\theta}) := \frac{\pi_g f_g^{ranef}(\mathbf{b}_i^g,\boldsymbol{\theta}^g)}{\sum_{\overline{g}=0}^{G-1}\pi_{\overline{g}} f_{\overline{g}}^{ranef}(\mathbf{b}_i^{\overline{g}},\boldsymbol{\theta}^{\overline{g}})}, \tag{3.12}$$

$$\mathcal{P}_{g,i}^{cond}(\mathbf{b}_i^0,...,\mathbf{b}_i^{G-1},\boldsymbol{\psi},\boldsymbol{\theta}) := \frac{\pi_g f_g^{cond}(\mathbf{y}_{i,1},...,\mathbf{y}_{i,R}|\mathbf{b}_i^g;\boldsymbol{\psi}^g)}{\sum_{\overline{g}=0}^{G-1}\pi_{\overline{g}} f_{\overline{g}}^{cond}(\mathbf{y}_{i,1},...,\mathbf{y}_{i,R}|\mathbf{b}_i^{\overline{g}};\boldsymbol{\psi}^{\overline{g}})}. \tag{3.13}$$

There are three different prediction approaches introduced. In chapter 5, these predictions are applied and the data set that is used in this thesis is classified using these approaches. The data set and software which is utilized in this thesis are introduced in the next chapter.

4. DATA AND SOFTWARE

4.1 Data

The data in this thesis contains longitudinal children's anthropometric measurements, height, and weight. These measurements have been collected from 4223 Finnish children who live in the Pirkanmaa area in Finland. Most of the children lived in Tampere and the rest of the children were from other municipalities near Tampere. The child's living area, urban or rural, is included in the data. The child's gender is included in the data as well and there were 2269 boys (53,7%) and 1954 girls (46,3%). Children were born in 1974, 1981, 1991, 1995, and 2001. The first four cohorts were measured eight times: at birth and when they were 6 months, 1, 2, 5, 7, 12, and 15 years old. Only children who were born in 2001 were measured only up to 11 years of age. This data set is investigated already in other studies. For example, Vuorela [24] and Nummi et al.[8] have investigated this same data set but the research questions were different.

4.1.1 Pre-processing

In this thesis, data has been split by gender because girls and boys grow up at different rates because girls and boys have different hormones. Puberty comes usually earlier for girls than for boys. It is interesting to investigate if there are differences between girls' and boys' predictability. Values of weight and height are used to calculate BMI. In the training data, BMI information is used from 1 up to 15 years of age to get the mixed models for both groups, overweight and normal weight, and the models are then exploited to predict those children in the test set who are at risk of being overweight at 15 years of age using their historical data which are collected from 1 year of age up to 7 years of age. Data includes measurements also at birth and at 6 months but these are omitted because the BMI curve that is shown in figure 4.2 has a big increase in the first year of a child's life and after that, the BMI curve fits better to second degree polynomial function. Since the whole curve is difficult to model by a simple polynomial function, the two first measurements are left out. The training set is utilized to get the model that is then applied to the test set. The test set uses data from 1 year of age up to 7 years of age to classify children who are at risk of being overweight at 15 years of age which means that their BMI is equal to or more than 25. Group 0 denotes normal weight and group 1 denotes overweight.

Table 4.1. *Count of children*

	Boys	Girls	All
All	2269	1954	4223
After pre-processing	1865	1592	3457
Training set	1305	1114	2419
Test set	560	478	1038
Overweight at 15 years of age in the training set (percentage)	137 (10.8%)	131 (12.1%)	268 (11.4%)
Normal weight at 15 years of age in the training set (percentage)	1128 (89.2%)	951 (87.9%)	2079 (88.6%)

There are some erroneous values in the data set because there is more than one measurement for one child for a particular age. There should be only one data point for each child each time. It has been modified so that the first data point is used for each child and if there is more than one, the rest of the data points are left out. Some children have no data for the age of 15 years. This thesis leaves out these children because there is not this actual value available and therefore comparing prediction results to the actual value is not possible with them.

Because this thesis predicts those who are at risk of being overweight at 15 years of age and children who are born in 2001 do not have data up to 11 years of age, this last cohort is omitted. After leaving out children born in 2001, data includes 1865 boys and 1592 girls. The split data sets, which include only boys in another data set and only girls in another data set and do not include children who are born in 2001, are then randomly raffled into training and test sets, 70% as a training set and the remaining 30% as a test set.

This thesis uses one marker to classify children into groups of normal weight and overweight. BMI was first used as a marker in this thesis but Nummi et al. [8] have used the transformation of BMI in their work and have shown that the results are better with transformation. Results from using BMI and inverse transformation of BMI were compared and it turned out that the results of using inverse transformation of BMI as a marker were better. Therefore the inverse transformation of BMI is utilized in this thesis.

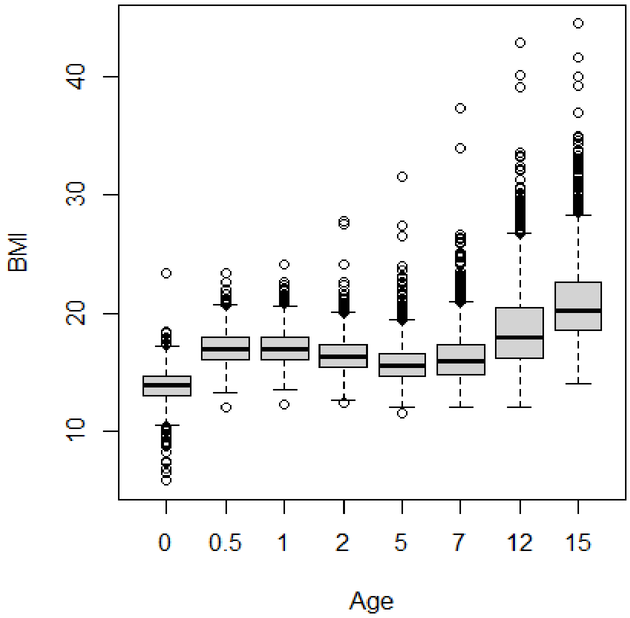


Figure 4.1. *Girls' mean development of BMI in different ages described by Box-Wisher plot at each measurement point*

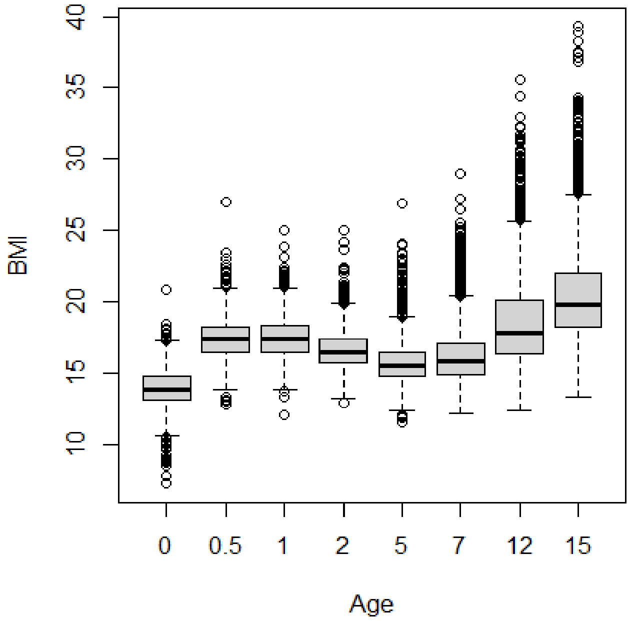


Figure 4.2. *Boys' mean development of BMI in different ages described by Box-Wisher plot at each measurement point*

4.2 R-package

This thesis uses package mixAK. It is an R-package for multivariate normal mixture models and mixtures of generalized linear mixed models including model-based clustering. In the first place it was proposed by Komarek [25] and after that, the package has been extended [26]. In the first place, mixAK was implemented for Bayesian estimation of mixtures of multivariate generalized linear mixed models. This package is capable of longitudinal data and classification of subjects into groups based on longitudinal data [26].

The package mixAK is used in this thesis and models are estimated using GLMM_MCMC -function. Mixed models will have the same structure in both groups which is introduced in chapter 5.1. The posterior inference for both data sets is based on 10 000 iterations of 1:10 thinned MCMC after a burn-in period of 5 000 iterations. GLMM_MCMC -function takes parameters as shown in the example in 5.1.1 as y is a vector or matrix with markers, x is a matrix with covariates for fixed effects, and z is a matrix with covariates for random effects.

Discrimination is done by using the GLMM_longitDA function which exploits fitted models GLMM_MCMC -function returns. GLMM_longitDA -function returns a list of probabilities $\mathcal{P}_{i,g}$ to include groups G using three different prediction approaches that were introduced in chapter 3.5, marginal, conditional, and random effect predictions. It gives probabilities for each child for every time point when a child has been measured. GLMM_longitDA -function is for actual discrimination and it takes the models that were implemented in GLMM_MCMC -function as parameters. The example of utilization is introduced in chapter 5.1.1.

5. RESULTS

This thesis aims to investigate how well can predict children who are at risk to become overweight at 15 years of age using their historical BMI data from one to seven years of age, t = (1, 2, 5, 7). The prediction uses MLMMs that were fitted to the training data. As described in the section 4.1.1, training data includes 70% of the whole data. As it can be seen in the table 4.1 training data contains 1305 boys and 1114 girls. Of these children, 137 were overweight at 15 years of age in the boys' group, and 131 were overweight in the girls' group when the limit of overweight was BMI >= 25 as it was described in the section 2. This thesis considers one longitudinal marker (r = 1) to classify children into two groups, normal weight and overweight (g = 0, 1). Nummi et al. [8] have shown that BMI distribution becomes more right-skewed and they have used transformation of BMI in their work to achieve normality. They have applied an inverse transformation in their work. In this thesis, we experimented both with and without transformation, and it turned out that the results were better with transformation. Thus, this thesis utilizes the inverse transformation to BMI and uses inversely transformed BMI as the marker. The data is used from one year of age because BMI increases rapidly from birth up to one year of age, and after that BMI curve resembles a second-degree polynomial function as it was described in the section 4.1.1.

5.1 Modelling

Mixed models have the same structure in both groups, although this was not required with this approach. For the marker and both groups, we use age and square of age in a matrix with covariates for fixed effects, $\mathbf{X}$. Matrix for random effects, $\mathbf{Z}$, includes a child's birth cohort and birth weight. Intercept includes in the model, and it is assumed to be random. The model for the ith unit can be represented as

$$\mathbf{Y}_i = \begin{bmatrix} 1 \\ Age \\ Age^2 \end{bmatrix} \boldsymbol{\alpha} + \begin{bmatrix} 1 \\ Birth\ weight \\ Birth\ year \end{bmatrix} \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{5.1}$$

where $\boldsymbol{\alpha}$ is a vector of fixed effects for inverse BMI and $\mathbf{b}$ is a vector of individual random effects for inverse BMI and ϵ_i is a vector of random errors for inverse BMI on the ith unit. The R-package mixAK does not allow to use same variables for the fixed effects $\mathbf{X}$, and random effects $\mathbf{Z}$ in GLMM_MCMC -function. That is described in the article by Komarek and Komarkova [26]. They represent those model parameters would be unidentified in the scenario if $\mathbf{X}$ and $\mathbf{Z}$ use the same variables and the overall mean would be estimated twice. If it would be possible to choose the same variables, we would have chosen age and squared age for random effects as well.

Individual distributions that are combined in the form of the mixture distributions are called mixture components K. Nummi et al. [8] have done trajectory analysis on this dataset, and they have figured out which number of trajectories works best when overweight is a target of an investigation. It turned out that four trajectories pose the most natural interpretations. Therefore, also in this thesis, we take four mixture components, K = 4. The MGLMMs in each group are based on 10000 iterations of 1:10 thinned MCMC after a burn-in period of 5000 iterations. The models are fitted using the R package mixAK and the GLMM_MCMC function, as reported in the sections 4.2 and 5.1.1.

Using these estimated mixed models for both groups, discrimination of test data is done by the GLMM_longitDA function. Three different discrimination approaches are applied to test data, marginal, conditional, and random effects. These three different prediction approaches were introduced in the section 3.5. Probabilities $(p^1, .., p^n)$ be at risk to become overweight at 15 years of age are computed for each child, each discrimination approach, and each time point. It means that for each child, there is more than one predicted value because there are predicted is the child at risk to become overweight at 15 years of age when the child is 1, 2, 5, and 7 years of age. This thesis considers every time point separately and compares them. Therefore it gives a comparison of is it whether or not easier to predict overweight when the child is older.

5.1.1 The use of R

Underneath are the functions that are used in this thesis. First, models are created by the GLMM_MCMC function for both the group of normal weight and the group of overweight, and using these estimated mixed models discrimination is done by the GLMM_longitDA function.

```
library(mixAK)

mod0 <- GLMM_MCMC(
    y=groupNormalweightGirls["inverseBMI"],
    dist = "gaussian",
    id = groupNormalweightGirls[,"child"],
    x = list(groupNormalweightGirls[,c("ttime", "ageSquare")]),
    z = list(groupNormalweightGirls[,c("year", "bweight")]),
    random.intercept = TRUE,
    prior.b = list(Kmax=4),
    nMCMC = c(burn=5000, keep=10000, thin=10, info=500)
    )

mod1 <- GLMM_MCMC(
    y=groupOverweightGirls["inverseBMI"],
    dist = "gaussian",
    id = groupOverweightGirls[,"child"],
    x = list(groupOverweightGirls[,c("ttime", "ageSquare")]),
    z = list( groupOverweightGirls[,c("year", "bweight")]),
    random.intercept = TRUE,
    prior.b = list(Kmax=4),
    nMCMC = c(burn=5000, keep=10000, thin=10, info=500),
    PED = FALSE
    )

clust <- GLMM_longitDA(
    mod = list(mod0, mod1),
    w.prior = c(0.9, 0.1),
    y = TestGirls["bmi"],
    id = TestGirls[, "child"],
    time = TestGirls[, "ttime"],
    x = list(TestGirls[,c("ttime", "ageSquare")]),
    z = list(TestGirls[,c("year", "bweight")]),
    xz.common = TRUE
    )
```

The same procedure is repeated for the group of boys also.

5.2 Example of one child's prediction

We select one random child as an example to see how the classification works. This randomly selected child is a girl born in 1974 and her birth weight is 3.6 kg. There are initial data in the table 5.1. The symbol 2 means a girl for sex and for residence, it signifies rural municipality. The data includes knowledge of whether a child is from the city of Tampere or from three rural municipalities in the same region [8]. Nevertheless, the area was not the target of investigation in this thesis.

Table 5.1 shows us that the child is not overweight at 15 years of age as her BMI is under

Table 5.1. *Initial data of one example child*

Child	Birth year	Weight (kg)	Height (cm)	Sex	Residence	Age	BMI
1027	1974	3.6	50.0	2	2	0.0	14.4000
1027	1974	6.6	64.0	2	2	0.5	16.1133
1027	1974	10.7	75.5	2	2	1.0	18.7711
1027	1974	15.0	89.0	2	2	2.0	18.9370
1027	1974	24.5	120.0	2	2	5.0	17.0139
1027	1974	27.0	126.0	2	2	7.0	17.0068
1027	1974	40.0	148.0	2	2	12.0	18.2615
1027	1974	55.5	164.5	2	2	15.0	20.5098

25. In the tables 5.2, 5.3, 5.4 are represented the predicted values from marginal, conditional, and random effect predictions. There are reported results gotten by GLMM_MCMC -function. Thus, there is the probability to be overweight at 15 years of age, and also the prediction class, where 0 is denoting the group of normal weight and 1 means the group of overweight. The values are predicted when the child is 1, 2, 5, and 7 years of age because the test set includes data from that range, and the goal was to predict if the child is at risk to be overweight when the child is 15 years old.

Table 5.2. *Results of marginal prediction for the example child*

Child	Age	Probability of overweight	Prediction class
1027	1	0.1047	0
1027	2	0.1312	0
1027	5	0.0846	0
1027	7	0.0348	0

Table 5.3. *Results of conditional prediction for the example child*

Child	Age	Probability of overweight	Prediction class
1027	1	0.0873	0
1027	2	0.0963	0
1027	5	0.0519	0
1027	7	0.0181	0

Marginal and conditional prediction approaches forecast correctly to prediction class 0 at every time point, the child is not going to be overweight. Merely, the random effect prediction approach forecasts overweight when the child is one, two, or five years old. That changes when the child is seven years old and at that time point this approach also predicts that at 15 years of age, the child is normal weight. The probabilities change a

Table 5.4. *Results of random effect prediction for the example child*

Child	Age	Probability of overweight	Prediction class
1027	1	0.9545	1
1027	2	0.9985	1
1027	5	0.9898	1
1027	7	0.2035	0

lot, for the first probability to be in the group overweight was close to 100% and at the last time point, it was 20%.

5.3 Methods to evaluate results

There are statistics for three different prediction approaches in the tables 5.5 and 5.6. TP is an abbreviation for true positives, FP for false positives, TN means true negatives and FP means false negatives. Further, TPR means the true positive rate also known as sensitivity, and TNR means the true negative rate also known as specificity. Sensitivity and specificity are defined as

$$\text{Sensitivity} = \text{TPR} = \frac{\text{True positives}}{\text{True positives + False negatives}}, \tag{5.2}$$

$$\text{Specificity} = \text{TNR} = \frac{\text{True negatives}}{\text{True negatives + False positives}}. \tag{5.3}$$

As can be seen in the tables 5.5 and 5.6, random effects prediction works worse than marginal and conditional predictions because accuracy is low. Accuracy is defined as

$$\text{Accuracy} = \frac{\text{True positives + True negatives}}{\text{True positives + True negatives + False positives + False negatives}}. \tag{5.4}$$

Table 5.5. *Values of TP, FP, TN, FN, TPR, TNR and accuracy for girls*

	TP	FP	TN	FN	TPR	TNR	Accuracy
Marginal	27	11	1701	197	0.1205	0.9936	0.8926
Conditional	12	4	1708	212	0.0536	0.9977	0.8884
Random effect	140	1295	417	84	0.6250	0.2436	0.2877

Table 5.6. *Values of TP, FP, TN, FN, TPR, TNR and accuracy for boys*

	TP	FP	TN	FN	TPR	TNR	Accuracy
Marginal	42	30	1974	224	0.1579	0.9850	0.8881
Conditional	49	33	1971	217	0.1842	0.9835	0.8899
Random effect	176	1743	261	90	0.6617	0.1302	0.1925

5.3.1 ROC curve

The receiver operating characteristic (ROC) curve shows the performance of the diagnostics test where is binary classifier [27]. This thesis uses it to evaluate classification models and how well classification works. ROC curve is advisable for visualizing classifiers' performance [28]. It is composed of a true positive rate on the y-axis and a false positive rate on the x-axis. True positive rate is a proportion of samples correctly classified to be positive out of all positive samples [28], as described in formula 5.2. In this thesis, it means the proportion of children that were classified correctly in the group of overweight out of all children who are classified as overweight. Accordingly, a false positive rate which is also known as 1 - specificity, is a proportion of samples that were incorrectly classified positive out of all negative samples [28]. It is shown in the formula 5.5. Thus, it means the proportion of children that were classified incorrectly to the group of overweight out of all children who are normal weight.

$$1 - \text{Specificity} = \text{FPR} = \frac{\text{False positives}}{\text{False positives + True negatives}} \tag{5.5}$$

Sensitivity, specificity, and ROC curves are commonly used in medical research but also in machine learning and data mining research [28].

In the figures, 5.1 and 5.2 are plotted three different ROC curves of predictions that the GLMM_longitDA -function gives for girls and boys and for each time point separately. Because the true positive rate is in the y-axis ROC curve can be interpreted as the nearer line is for the y-axis, the better the performance is. For this is very common to calculate an area under the ROC curve (AUC) [28].

5.3.2 Area under the ROC curve

To interpret the ROC curve and get the overall results of the diagnostics test's performance area under the ROC curve (AUC) is calculated [27]. AUC tells how well a model can classify cases. If the AUC equals 1 it is the most ideal situation and the classifier classifies all cases correctly. Vice versa, if the AUC equals 0, it signifies that the classifier predicts all the cases backward. On the other hand, if the AUC equals 0.5 it is the worst

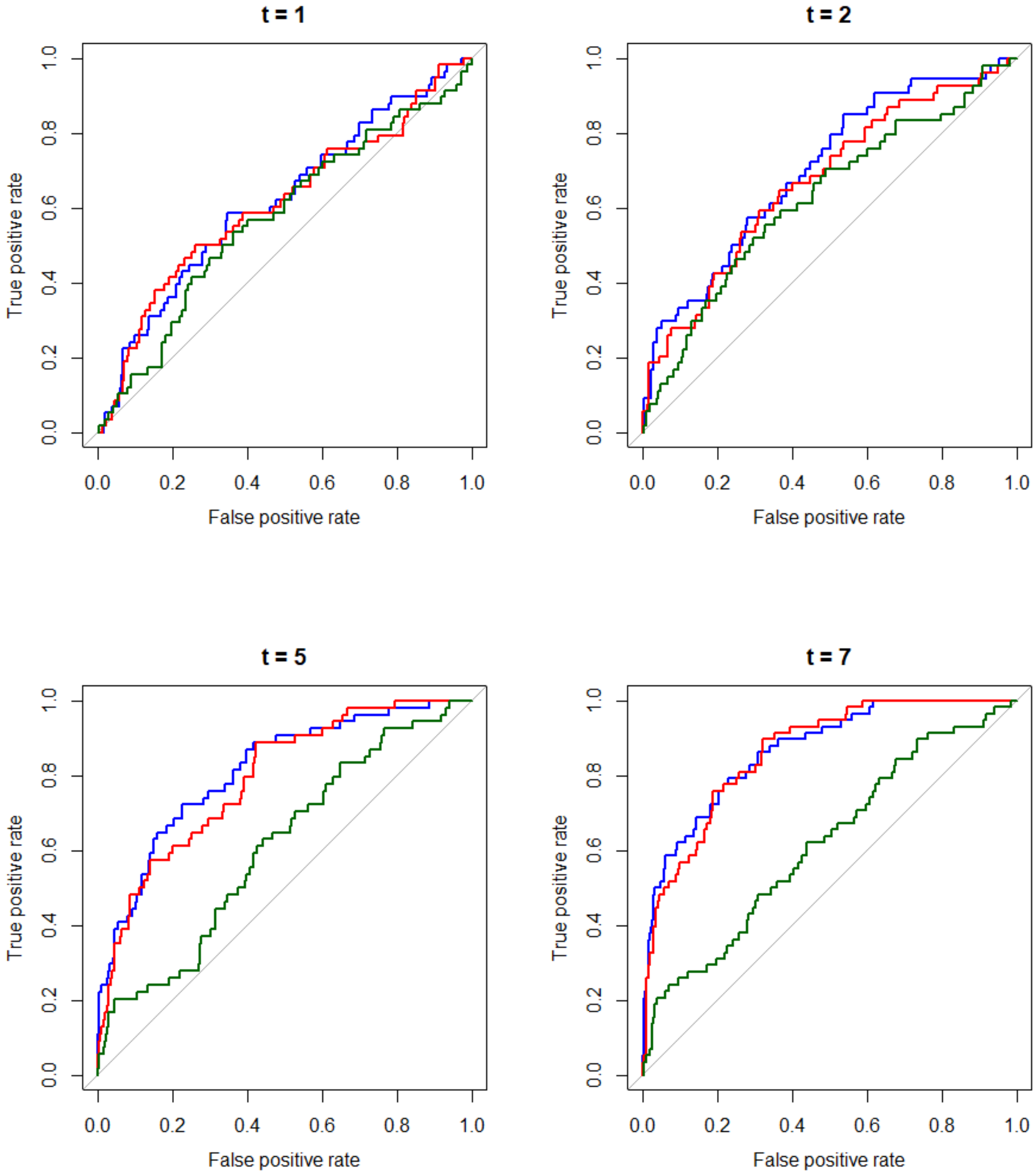


Figure 5.1. *Girls' ROC curves. The blue lines signify the results of a marginal prediction, the red lines signify the results of a conditional prediction and the green lines signify the results of a random effect prediction.*

situation and it can not be known at all whether could classifier classify cases correctly or not.

5.4 Comparison of prediction methods

The figures 5.3 and 5.4 show probabilities of being overweight at 15 years of age, and there are illustrated all three different prediction approaches. In all the figures, the time

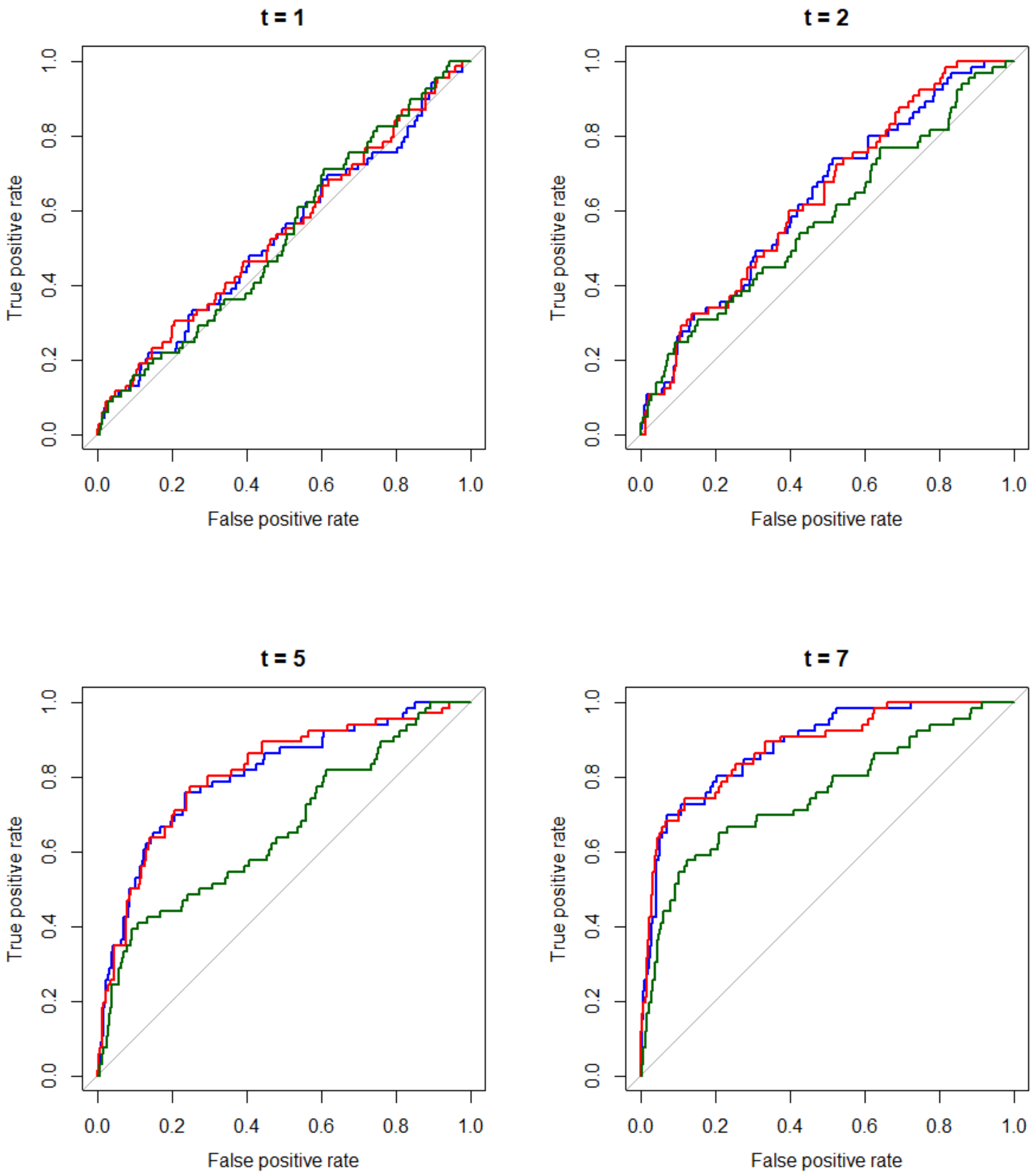


Figure 5.2. *Boys' ROC curves. The blue lines signify the results of a marginal prediction, the red lines signify the results of a conditional prediction and the green lines signify the results of a random effect prediction.*

is on the x-axis and one line represents one child. Every probability is calculated using the longitudinal data that is available at that time. Therefore it is possible to see how the probability varies when a child becomes older and there are more longitudinal data available for the child. If the probability is higher than 0.7 a child is classified in the group of overweight. In marginal and conditional predictions there are none of the children whose probability is higher than 0.7 in the first time point both in the girls' and boys' groups. It is interesting why no one is predicted to be in the group of overweight based on BMI at

1 year of age. Also, there are only a few children whose probability is higher than 0.7 in the second time measurement. Over time there are more children who are predicted to be overweight. As the figures 5.3 and 5.4 show, the random effect prediction gives totally different results. There are much higher probabilities to be overweight in random effect prediction and the probabilities in both groups seem to vary much and there are no obvious trends to be seen.

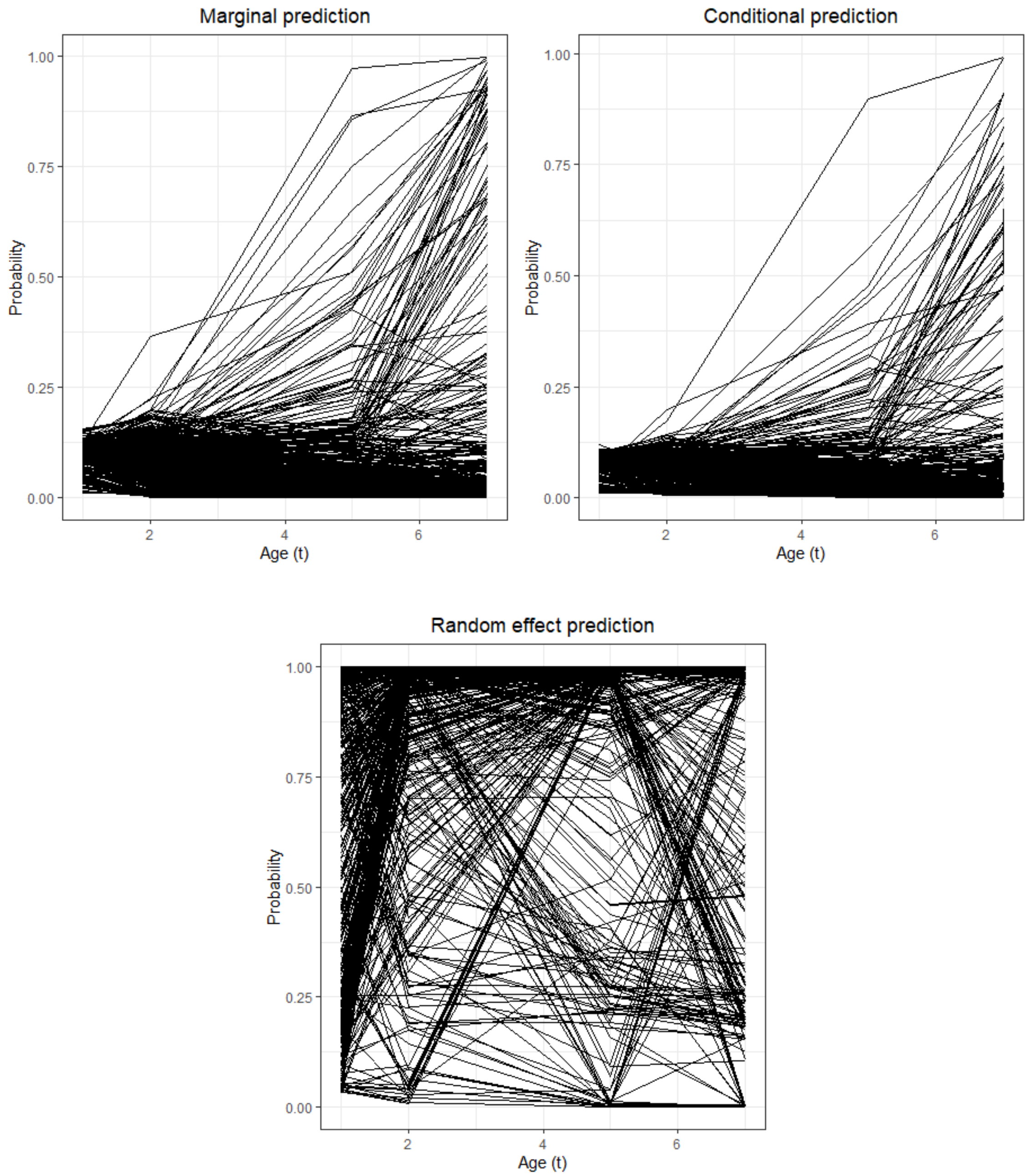
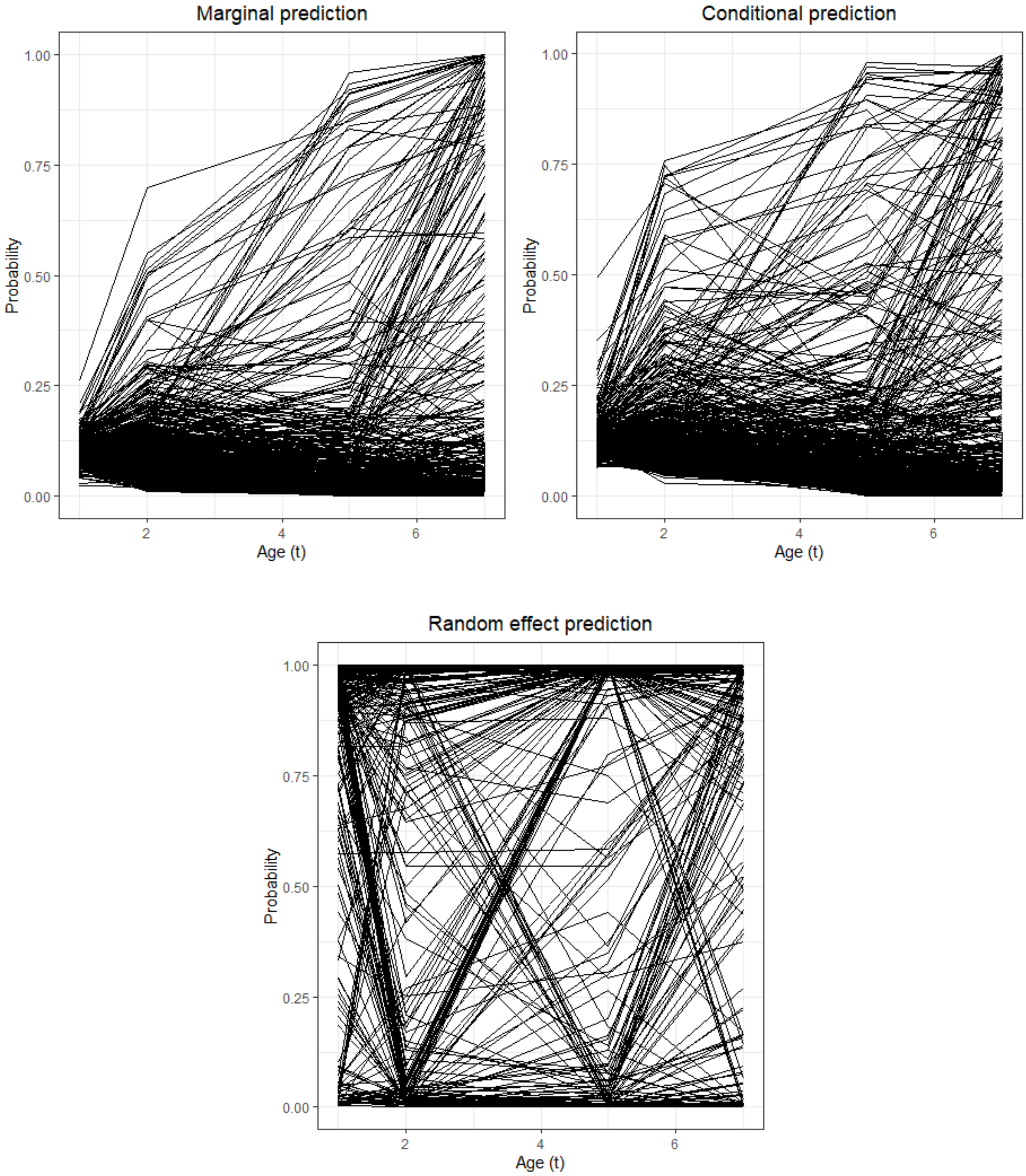


Figure 5.3. *Girls' probabilities of being overweight at 15 years of age*

It can be seen in the tables 5.5 and 5.6 that specificity is better than sensitivity in marginal and conditional predictions but the opposite way in random effect prediction. The marginal and conditional predictions forecast hardly any child to become overweight so it is under-



Marginal prediction

Conditional prediction

Random effect prediction

Figure 5.4. *Boys' probabilities of being overweight at 15 years of age*

standable that sensitivity is weak in these predictions. There were only 10.8% overweight children in the boys' training set and 12.1% in the girls' training set. That might affect how easy is to predict which children are at risk to become overweight, in the other words predict positives, and how easy to predict negatives because there are different amounts of positives and negatives in the training data. There were more children with normal weight in the training data. The random effect prediction forecasts much higher probabilities to be overweight, and that is a reason why there are more true positive values. Hence, sensitivity is higher in random effect prediction than in marginal and conditional prediction approaches. Specificity in random effect prediction is poor in both groups because that approach predicts plenty of false positives.

The random effect seems to be the worst prediction approach regardless of whether the target is girls or boys if regards the figures 5.1 and 5.2. Marginal and conditional predictions give results along the same lines, and there is no big difference between them. The random effect is going near the middle of the ROC curve, and the AUC values are along the side, especially in the girls' group. The AUC values are reported in the tables 5.7 and 5.8. AUC is nearby 0.5 when children are 1 year old in all prediction approaches. As we can see, the AUC value is bigger over time, but in the girls' group, the trend is not so obvious as in the boys' group in the random effect prediction. In the marginal and conditional predictions improvement of the AUC values are remarkable over time. It is reasonable that when children are older it is easier to predict whether are they going to be overweight or not later in life. Considering ROC curves and AUC values, we can notice that in the first two time points, when children were 1 and 2 years of age, none of the prediction approaches worked properly either in the girls' or boys' groups. It is evident that there is a remarkable improvement in the performance of all predictions in the boys' group over time and in the girls' group with marginal and conditional predictions. In the first two data points predictions are poor but in the last time point, results are much better.

Table 5.7. *Girls' area under the ROC curve (AUC)*

	Marginal	Conditional	Random effect
t = 1	0.6169	0.6099	0.5717
t = 2	0.7010	0.6698	0.6250
t = 5	0.8125	0.7910	0.6079
t = 7	0.8669	0.8626	0.6174

Table 5.8. *Boys' area under the ROC curve (AUC)*

	Marginal	Conditional	Random effect
t = 1	0.5309	0.5395	0.5305
t = 2	0.6305	0.6342	0.5784
t = 5	0.8045	0.8080	0.6557
t = 7	0.8815	0.8780	0.7496

In the random effect prediction, the focus is on the subject-specific evolution of longitudinal profiles and it is grounded on the distributions of the individual random effects. The matrix for random effects includes a child's birth weight and birth cohort in the model that was utilized in this thesis. Because the results are weak with this prediction approach it would mean that birth weight and birth cohort do not have much influence on BMI later in life with these children. This is also contributed by the fact that in the marginal prediction approach, the random effects are integrated out, as was described in the section 3.5, and the marginal prediction outperform the random effect prediction. This is exceedingly a surprising result because the hypothesis was that random effect prediction works well and the individual random effects have an influence on possible overweight in the future. One reason might be an issue that it was not possible to select the same variables in the matrices for fixed effects $\mathbf{X}$, and random effects $\mathbf{Z}$ in GLMM_MCMC -function, as it was told in 5.1. If it would have been possible, we have chosen a child's age and square of age for the matrix for random effects, and that would have improved the capability to classify children into the right classes.

In summary, it seems that marginal and conditional predictions work better than random effect predictions. Results do not differ much whether the target is girls or boys, although, in the boys' group, the improvement of predictability was more notable than in the girls' group, especially with random effect prediction. One visible and also intuitive result was that the older child is, the easier is to predict possible overweight.

6. CONCLUSION

This thesis used discriminant analysis to predict whether children are at a risk to become overweight later in life based on their longitudinal historical data. The aim was to create a linear mixed model with a normal mixture in the random effects distribution based on the training data. Using these created models the purpose was to classify children into two groups using Bayes' theorem and Markov Chain Monte Carlo method. Three different prediction approaches were utilized in this thesis. Marginal, conditional, and random effect predictions are commonly used in longitudinal discriminant analysis based on mixed models, and the main idea is to compare group probabilities that are based on historical data and then set a subject into the group whose probability is the biggest. The model that was utilized and the predictions were based on includes one marker, inverse BMI. We used also BMI as a marker but it turned out that transformation produced better results and therefore inverse BMI was applied, as it was applied in the study by Nummi et. al [8] which was done using this same data set. In this thesis, a child's age and square of age were fixed effects and random effects were the child's birth weight and birth year in the model.

The analysis is based on the data set which includes Finnish children's anthropometric measurements from birth up to 15 years of age, and it was collected in the Pirkanmaa area in Finland. There were children from the city and rural municipality, and although the area was not the target of the investigation, it is important to become aware that data represent most parts of Finland. In this thesis, data were separated by gender and the results were considered separately by girls and boys. There were no observable differences between girls and boys. The only notice was that when boys were one or two years of age it was not possible at all to predict whether they were at risk to become overweight or not. Regarding girls that were also very difficult but it seems that for girls it is slightly easier than for boys. Girls and boys grow up at different rates and also puberty comes earlier for girls than boys usually. It can be one reason why it is a little bit easier to predict if they are overweight at 15 years of age. The data was separated by gender because of this difference between girls' and boys' growth. The reason why to predict overweight at 15 years of age was the simply that data set contains values up to 15 years of age and then it was possible to compare predicted values to actual values. Also, the improvement of predictability was a little bit better with boys than girls over time.

Three different prediction approaches were used and there were differences between them. Marginal and conditional prediction approaches accomplished better results than random effect prediction which was surprising. One of the limitations was that the same variables were not able to use both fixed effect and random effect in the R package that was applied in this thesis, and it may have impacted the quality of the model and the model's predictability. The results improved over time in both the girls' and the boys' groups. It is conceivable that the older a child becomes the easier is to see whether the child will be overweight or not.

This data set was investigated in earlier studies also but the target of the investigation was different in this thesis. Nummi et. al [8] identified trajectories in this data set and based on their work this thesis used four mixture components for modeling. Nummi et. al [8] have also used inverse transformation of BMI in their work and inverse transformation of BMI was utilized also in this thesis as the marker.

This thesis considered gender but not residence. For future studies, it would be interesting to investigate more about the impact on the area where a child lives, in a city or rural area. Also, it would be interesting to investigate more how birth year affects BMI and if there are differences between children who have born in the 20th and 21st centuries.

REFERENCES

[1] *Report of the Commission on Ending Childhood Obesity: implementation plan: executive summary*. Tech. rep. World Health Organization, 2017.

[2] World Obesity Federation: Lobstein, T. and Brinsden, H. *Atlas of Childhood Obesity*. Version 1. 2019. URL: https://www.worldobesity.org/membersarea/global-atlas-on-childhood-obesity.

[3] Hauerslev, M., Narang, T., Gray, N., Samuels, T. A. and Bhutta, Z. A. Childhood obesity on the rise during COVID-19: A request for global leaders to change the trajectory. *Obesity* 30.2 (2022), pp. 288–291.

[4] FAO. Percentage of people worldwide who were overweight as of 2019, by age. *Statista* (2019). URL: https://www-statista-com.libproxy.tuni.fi/statistics/1065605/prevalence-overweight-people-worldwide-by-age/?locale=en.

[5] Sahoo, K., Sahoo, B., Choudhury, A. K., Sofi, N. Y., Kumar, R. and Bhadoria, A. S. Childhood obesity: causes and consequences. *Journal of family medicine and primary care* 4.2 (2015), p. 187.

[6] Jordan, M. I. and Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* 349.6245 (2015), pp. 255–260.

[7] Dick, S. Artificial Intelligence. *Harvard Data Science Review* 1.1 (2019). URL: https://hdsr.duqduq.org/pub/0aytgrau.

[8] Nummi, T., Hakanen, T., Lipiäinen, L., Harjunmaa, U., Salo, M. K., Saha, M.-T. and Vuorela, N. A trajectory analysis of body mass index for Finnish children. *Journal of Applied Statistics* 41.7 (2014), pp. 1422–1435.

[9] Šedová, L., Tóthová, V., Olišarová, V., Adámková, V., Bártlová, S., Dolák, F., Kajanová, A., Mauritzová, I., Nováková, D. and Prokešová, R. Evaluation of selected indicators of overweight and obesity of Roma minority in the region of South Bohemia. *Neuro Endocrinol Lett* 36.Suppl 2 (2015), pp. 35–42.

[10] Karchynskaya, V., Kopcakova, J., Klein, D., Gába, A., Madarasova-Geckova, A., Dijk, J. P. van, Winter, A. F. de and Reijneveld, S. A. Is BMI a valid indicator of overweight and obesity for adolescents?: *International journal of environmental research and public health* 17.13 (2020), p. 4815.

[11] Rothman, K. J. BMI-related errors in the measurement of obesity. *International journal of obesity* 32.3 (2008), S56–S59.

[12] Learned-Miller, E. G. Introduction to supervised learning. *I: Department of Computer Science, University of Massachusetts* (2014), p. 3.

[13] Nasteski, V. An overview of the supervised machine learning methods. *Horizons. b* 4 (2017), pp. 51–62.

[14] Gnanadesikan, R. *Discriminant analysis and clustering*. National Academies Press, 1988.

[15] Hughes, D. M., Komárek, A., Czanner, G. and Garcia-Finana, M. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 27.7 (2018), pp. 2060–2080.

[16] Komárek, A., Hansen, B. E., Kuiper, E. M., Buuren, H. R. van and Lesaffre, E. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in medicine* 29.30 (2010), pp. 3267–3283.

[17] Gabrio, A., Plumpton, C., Banerjee, S. and Leurent, B. Linear mixed models to handle missing at random data in trial-based economic evaluations. *Health Economics* 31.6 (2022), pp. 1276–1287.

[18] Winter, B. A very basic tutorial for performing linear mixed effects analyses. *arXiv preprint arXiv:1308.5499* (2013).

[19] Berrar, D. Bayes' theorem and naive Bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* 403 (2018), p. 412.

[20] Kalos, M. H. and Whitlock, P. A. *Monte carlo methods*. John Wiley & Sons, 2009.

[21] Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. *Handbook of markov chain monte carlo*. CRC press, 2011.

[22] Carlo, C. M. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB* 581.540 (2004), p. 3.

[23] Morrell, C. H., Brant, L. J. and Sheng, S. Comparing approaches for predicting prostate cancer from longitudinal data. *2007 Proceedings of the American statistical association* (2007), pp. 127–133.

[24] Vuorela, N. *Body Mass Index, Overweight and Obesity Among Children in Finland-A Retrospective Epidemilogical Study in Pirkanmaa District Spanning Over Four Decades*. Tampere University Press, 2011.

[25] Komárek, A. A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Computational Statistics & Data Analysis* 53.12 (2009), pp. 3932–3947.

[26] Komárek, A. and Komárková, L. Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software* 59 (2014), pp. 1–38.

[27] Park, S. H., Goo, J. M. and Jo, C.-H. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean journal of radiology* 5.1 (2004), pp. 11–18.

[28] Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* 27.8 (2006), pp. 861–874.