BACHELOR

Exploration on Risk Profiles for Obesity/Overweight

Drugă-Tache, Robert

Award date:
2024

Exploration on Risk Profiles for Obesity/Overweight

Robert Druga-Tache*
Student Number: 1768468

*e-mail: r.druga.tache@student.tue.nl

CONTENTS

1 INTRODUCTION

This paper aims to contribute to the conversation about obesity among children by coming up with new insights regarding the issue. The following analysis is based on data collected within The Brabant Study [1].

1.1 Obesity Explained

The World Health Organization defines obesity as an "abnormal or excessive fat accumulation that presents a health risk"[2]. This condition is an effect of an energy imbalance between calories consumed and calories expended and can be influenced by various genetic, behavioral, and environmental factors, with a diagnosis of overweight and obesity given based on the Body Mass Index measure[2].

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height}^2(\text{m}^2)}$$

To diagnose a person as obese or overweight, the World Health Organization differentiates between age and gender in infants, children, and adolescents[2], as displayed in Table 1(compiled based on the definition of obesity offered in [2])

Group	BMI Category
Adults	Overweight: BMI $\geq$ 25 Obesity: BMI $\geq$ 30
Children under 5 years	Overweight: Weight-for-height $>$ 2 SD above WHO median Obesity: Weight-for-height $>$ 3 SD above WHO median
Children 5-19 years	Overweight: BMI-for-age $>$ 1 SD above WHO median Obesity: BMI-for-age $>$ 2 SD above WHO median

Table 1: BMI Categories for Defining Obesity by Age Group according to the WHO[2]

1.2 Consequences of obesity and overweight

According to NIH: The National Institute of Diabetes and Digestive and Kidney Diseases, obesity can lead to severe conditions like[3]:

- **Type 2 Diabetes and High Blood Glucose**: These conditions can lead to complications such as heart disease, stroke, kidney disease, eye problems, and nerve damage[3].
- **High Blood Pressure**: This increases the risk of heart attack, stroke, and kidney disease[3].
- **Heart Disease**: People who are overweight and obese are more likely to experience heart attack, heart failure, and abnormal heart rhythms[3].
- **Metabolic Syndrome**: A cluster of conditions that elevate the risk for heart disease, diabetes, and stroke[3].
- **Breathing Problems**: Obesity can lead to various respiratory issues[3].
- **Asthma**: Obesity is a known risk factor for developing asthma[3].
- **Fertility Problems**: There is an increased risk of infertility associated with obesity[3].
- **Gout**: This type of arthritis, characterized by joint pain and swelling, is caused by the buildup of uric acid crystals[3].
- **Gestational Diabetes**: Obesity increases the risk of developing diabetes during pregnancy[3].

These findings are also supported by additional research [4].

1.3 Obesity and overweight as a pandemic

There are a variety of factors that make obesity and overweight global threats. One of those is the prevalence of ultra-processed foods: people in the highest percentile of ultra-processed food consumption are at a higher risk of developing overweight and obesity[5].

In 2015, there were 107.7 million obese children and 603.7 million obese adults globally[4]. Recent trends have shown a 1.7 % rise in the prevalence of severe child obesity. Moreover, if only studies with a follow-up analysis of 20 years are included, a notable increase of 9.6 % can be noticed [6].

Solutions for tackling these issues exist. Cities with built environment characteristics like walkability, the presence of sidewalks or green spaces are a solution to tackle obesity and overweight among children[7]

1.4 Link Between Childhood Obesity and Adult Diseases

Childhood obesity and overweight lead to many issues during adulthood. Some studies [8] focus on the link between a person being overweight in childhood, and later developing type 2 diabetes, hypertension, or coronary heart disease in adulthood [8].

1.5 Public Policies for combating childhood obesity

According to [9], low socioeconomic status is an important factor in childhood and adolescent obesity. Thus, it can be said that tackling childhood obesity is intertwined with providing children at risk of poverty with key services for their development. To that end, the European Commission adopted in 2019 the European Child Guarantee to ensure basic needs for all children at risk of poverty in Europe: among these basic needs, there is guaranteed access to healthy nutrition [10].

Member States are asked within this plan to develop national action plans to achieve the goals within the proposed document. For example, one program that's part of the national action plan in the Netherlands is the Healthy School Strategy: it promotes healthy eating and lifestyles among children by integrating nutrition into education, setting policies for healthy snacks, and fostering a supportive school environment. The program aims for one-third of schools to adopt these practices by 2024, with the long-term goal being for all schools to have a Healthy School coordinator by 2040[11]. Additionally, free fruit and vegetables are provided to 3,000 primary schools for 20 weeks a year under the EU scheme.

1.6 The Brabant Study :

The Brabant study is a longitudinal study involving 4000 pregnant women. They are recruited at 8-10 weeks gestation from community midwife practices in South-East Brabant, Netherlands [1]. The study includes two types of data: medical data which include Thyroid function parameters, and thyroid peroxidase antibody, assessed at 12, 20, and 28 weeks of pregnancy[1], and questionnaires on demographic and obstetric features, lifestyle habits, psychological and social variables, and partner relationship quality completed by the participants within the study[1].

Thus, the objective of this paper is to leverage data collected in the study above and inform obesity and overweight in children, based on the mother's characteristics

1.7 Objectives of the analysis

Studies show that there is a significant link between maternal Obesity and child obesity[12]. That is why, the objective of this analysis is to shift the focus to inform children

overweight based on data from a mother's pregnancy. The Brabant study was conducted in such a way that it recorded, besides medical and demographic data, lots of psychological data, thus allowing for a complete analysis of the matter. A big emphasis will be also put on mental-health-related factors. The methodology involved in approaching this question will involve predictive or causal analysis. The main objective of this paper is to explore factors within the data leading to overweight and obesity and try to relate them to other studies.

1.8 Research Questions

This analysis aims to answer the following primary research question: **What are the risk factors leading to overweight/obese children, and how can obesity be prevented or mitigated through predictive or causal analysis for unborn children?**

The main research questions will be explored with the aid of two research sub-questions:

- *How does predictive modeling perform in identifying the risk of childhood overweight?*
- *What causal relationships exist between childhood overweight and maternal characteristics?*

2 Literature Research

In the study "Explorations on risk profiles for overweight and obesity in 9501 preschool-aged children" [13], preschool-aged children in China were examined to gather data on obesity and overweight, following WHO definitions [2] regarding obesity/overweight classifications. Methods in this study include stratified random sampling, group division into overweight/obesity and non-overweight groups, statistical comparisons, and regression analyses. I intend to adopt similar methods in my research, focusing on group division and between-group comparisons. I intend to adopt in my study approaches like division into treatment and control groups and between-group comparisons. The results of [13] suggest sleep duration, birth-weight, paternal BMI, maternal BMI, gestational weight gain (GWG), and maternal pre-pregnancy BMI are significantly associated with overweight and obesity in preschool-aged children. Maternal pre-pregnancy overweight and obesity are determined as risk factors for childhood overweight and obesity in [14] as well.

Given my study's focus on also revealing mental-related risk factors for obesity, the results of [15] indicate a negative correlation between maternal self-esteem and the likelihood of her children developing obesity.

The longitudinal nature of the data entails the research for longitudinal approaches to analyzing data. Common issues with these types of datasets are dealing with missing values and the decision on how to rank features, and which features to use in the model. In [16], the missing values issue is solved through Multivariate Imputation, using Chained Equations.

3 Methodology

This project's analysis consists of two main phases: Data Preparation and Causal/Predictive Analysis.

During the Data Preparation phase, we identified and engineered important features using both automated and empirical methods. The primary goal of this phase was to reduce the data's dimensionality to make the subsequent predictive modeling and causal analysis more explainable.

In the Causal/Predictive Analysis phase, several techniques were used to uncover relationships and predict outcomes. Specifically, logistic regression, XGBoost, and instrumental variable (IV) regression were used to analyze the data. Logistic regression helped with identifying significant predictors of childhood overweight. XGBoost was used as an improvement for predictive modeling accuracy. IV regression allowed for the exploration of causal relationships between maternal characteristics and childhood overweight/obesity. Together, these methods provided a comprehensive understanding of the factors contributing to childhood overweight and how they can be predicted or mitigated.

4 Data Preparation

4.1 About the Data

The data for this analysis is sourced from the Brabant Study, a large prospective pregnancy cohort study following 1328 women from 12 weeks of pregnancy until their child reaches five years of age[1]. This study offers two sets of data: one tracking mothers during pregnancy and another monitoring the weight and length of the children after birth. The initial dataset comprises 1092 variables, encompassing psychological, demographic, and medical factors. Hence, the subsequent phase of the analysis involves examining and refining each category of features individually. An important goal of this step is dimensionality reduction for the dataset. Dimensionality reduction for the feature space of a dataset involves either the removal of certain features or combining multiple features into a single variable (for example, creating a work variable for each period to assess the work impact on a participant). During this stage, an empirical analysis was also conducted with the aid of the study's code book as well as with plots displaying the distributions of missing features and the number of missing values.

4.2 Preliminary Data Cleaning

Initially, features with more than 50% missing values and text columns not deemed relevant during the empirical analysis were eliminated. Relevant text columns were deemed only the ones from which certain numerical values could be extracted because 90% of all participants in the study provided answers in Dutch(as shown in Figure 1). Moreover, handling Dutch text posed additional challenges. Typically, respondents input Dutch text when asked to describe a category, especially if they don't identify with predefined categories in the questionnaire. This often occurs when they select "other" as their response, this option requiring text input.

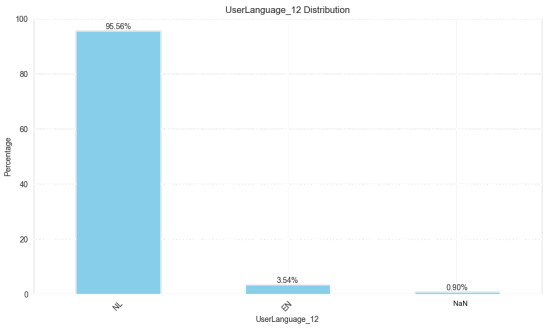

Figure 1: User Language Distribution

4.3 Feature Categorization

An important step in detecting the underlying patterns of the data is identifying and categorizing the most relevant features from the sourced dataset. 3 main feature categories were identified: demographic, medical, and psychological. Demographic features were recorded only at the baseline (12 weeks of pregnancy), while medical and psychological features were recorded throughout the pregnancy. Below, an

overview of the most important features within each category is presented.

- **Demographic features:**: Education, Marital Status, Planned Pregnancy (assess whether the participant's pregnancy was planned), Ethnicity, Breadwinner (primary financial provider of the family)
- **Medical features:** Obstetric variables, Family medical history on chronic diseases and mental health, Previous pregnancy/delivery problems, Chronic conditions, Hospitalization before birth, Maternal medical history, Medical appointments, Medication
- **Psychological features:**Psychological features are determined by medical history of previously diagnosed mental health issues as well as *results of questionnaires completed by each participant* at 12 weeks, 20 weeks, 28 weeks in their pregnancy, and at 8 weeks after giving birth.

4.4 Psychological data overview

In the Brabant study, psychological features were recorded with the help of several types of questionnaires given to each participant at different points in the study: 12 weeks of pregnancy, 20 weeks of pregnancy, 28 weeks of pregnancy, and 8 weeks after giving birth[1]. Thus, an individual's trait is assessed with the help of the total score of a questionnaire he completed measuring that trait. Table 2 presents the types of questionnaires conducted within the study and what they are assessing.

Table 2: Purpose of each type of questionnaire

Questionnaire Type	Measures
TPDS	Pregnancy Distress Level
EDS	Depression Level
SCL90	Anxiety Level
DAS	Relationship Satisfaction
IWPQ	Work Performance
PUQE	Morning Sickness Level
PPBS	Bonding with the Child
BFI_2S	Personality
DERS-16	Difficulties in Emotion Regulation
OBVL	Interaction with Newborn Baby
BSMAS	Social Media Addiction Level

4.4.1 Created Features

The following features were not derived from previous research but were created by summing groups of features found in the codebook of the Brabant Study data([1]), according to their grouping there. Thus, features assessing the same issue were combined.

- Interaction_newborn_baby:

$$\sum_{i\in\{1,2,5,6,9,10,13,14,17,18,21,22,25\}} \text{OBVL\_i\_8wPP\_r}$$

- Psychological_complaints_8wPP:

$$\sum_{c\in\{\text{SCL90\_columns\_8wPP}\}} \text{dataset}[c]$$

- Social_media_use:

$$12weeks: \frac{\text{BSMAS\_TOT\_12}}{\text{SMUfreq\_12\_r}+1}$$
$$20weeks: \frac{\text{BSMAS\_TOT\_20}}{\text{SMUfreq\_20\_r}+1}$$
$$28weeks: \frac{\text{BSMAS\_TOT\_28}}{\text{SMUfreq\_28\_r}+1}$$

- Difficulties in Emotion Regulation:

$$\text{DERS16\_Goals\_TOT\_20} + \text{DERS16\_Impulse\_TOT\_20} + \text{DERS16\_Strategies\_TOT\_20} + \text{DERS16\_Nonacceptance\_TOT\_20}$$

- Mindfulness:

$$\text{TFMQ\_SF\_Nonreacting\_TOT\_20} + \text{TFMQ\_SF\_Nonjudging\_TOT\_20} + + \text{TFMQ\_SF\_Awareness\_TOT\_20}$$

- Social Support:

$$\text{MSPSS\_Family\_TOT\_20} + \text{MSPSS\_Friends\_TOT\_20}$$

- Bonding:

$$\textbf{Bonding\_pre\_birth} = \text{PPBS\_TOT\_28}$$
$$\textbf{Bonding\_post\_birth} = \text{PPBS\_1\_8wPP\_r} + + \text{PPBS\_2\_8wPP\_r} + \text{PPBS\_3\_8wPP\_r}$$

- Work Mental Health:

$$\textbf{Work\_mental\_health\_12} = \text{VBBA\_Supervisor\_TOT\_12} + + \text{VBBA\_Colleagues\_TOT\_12} - \text{Work\_Engagement\_TOT\_12} - - \text{Work\_Burnout\_TOT\_12}$$
$$\textbf{Work\_mental\_health\_28} = \text{VBBA\_Supervisor\_TOT\_28} + + \text{VBBA\_Colleagues\_TOT\_28} - \text{Work\_Engagement\_TOT\_28} - \text{Work\_Burnout\_TOT\_28}$$

- Work Performance:

$$\textbf{Work\_performance\_12} = \text{IWPQ\_TP\_TOT\_12} + + \text{IWPQ\_CP\_TOT\_12}$$
$$\textbf{Work\_performance\_28} = \text{IWPQ\_TP\_TOT\_28} + + \text{IWPQ\_CP\_TOT\_28}$$

Table 3 provides an overview of the newly created numerical variables (derived from the questionnaire data) and the number of times they were recorded at different time periods throughout the pregnancy.

Table 3: Variables Recorded at Multiple Points

Variable category	Time Points measured
Social_media_use	**3**
Difficulties_emotion_regulation	1
Mindfulness	1
Social-support	1
Personality_BFI	1
Bonding_pre_birth	1
Bonding_post_birth	1
Type_D_personality	1
Pregnancy_distress	**3**
Partner_involvement	**3**
Morning_sickness	**3**
Depression	**3**
Anxiety	**3**
Relationship_satisfaction	**3**
Work_mental_health	2
Work_performance	2
Interaction_newborn_baby	1
Psychological_complaints_8wPP	1

4.4.2 Problems with questionnaire data

Certain types of questionnaires are not present throughout the entirety of the study. Some are specific to specific periods. As outlined before the periods in the dataset include 12 weeks of pregnancy, 20 weeks of pregnancy, 28 weeks of pregnancy, and 8 weeks postpartum. In the first 3 periods of pregnancy, standardized results of questionnaires, and individual questions in each are present. These questionnaires assess Nausea (PUQE questionnaire), Pregnancy distress (TPDS), Depression levels (EDS), Anxiety levels (SCL-90), Relationship Satisfaction (DAS), and Social Media Use(BSMAS/Frequency SMU) throughout the entirety of the pregnancy. Other important variables like the impact of work on mental health or personality are only measured at 12 weeks of pregnancy and 28 weeks of pregnancy. One important aspect that must not be forgotten is the absence of total scores for questionnaire data at 8 weeks after pregnancy (postpartum). In the 8 weeks post-pregnancy period, participants were administered a subset of questions from the standardized questionnaire. Below is an overview of the completeness of each type of questionnaire conducted during this period, expressed as the proportion of questions asked compared to the number of questions contained in a standardized questionnaire 2.

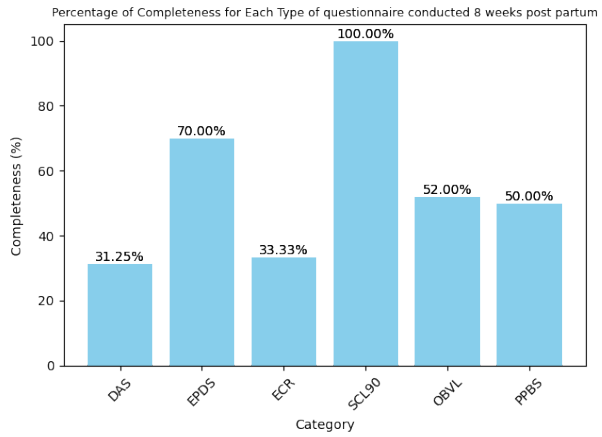

Figure 2: Completeness of each type of questionnaire conducted after pregnancy

As shown in Figure 2, the people conducting the study didn't ask all the questions from the standardized questionnaires. For example, only 33% of the standard questions contained within the Experiences in Close Relationships Scale Short Form were asked. This issue makes the mentioned questionnaires challenging to assess.

4.5 Medical data overview

The rest of the features in the dataset, after eliminating the psychological ones were deemed as medical. Medical features can be found in 2 datasets: one containing features recorded throughout the pregnancy of the mother, and another containing the recorded weights and heights of children up to 5 years of age.

4.5.1 *Feature Engineering for medical data*

After a careful analysis of the existing medical features in the data, new features were created with the aid of the textual descriptions women provided during the pregnancy.

- **Gynaecologist feature**: This feature records the week in which a participant was referred to a gynecologist.

  It was created by extracting the number of weeks from the text inputted by each participant at 20 and 28 weeks where she was asked to type in information about her gynecologist

- **Guided by obstetrician**: This binary feature was extracted using the same method as for the previously mentioned feature, and it records if the participant was guided by an obstetrician throughout the pregnancy

- **Nausea**: This feature records the number of weeks in the pregnancy up to which the participant experienced nausea-related symptoms.

  It was created by taking the maximum out of the extracted numbers in 3-time points in which questions about nausea were asked: 12 weeks, 20 weeks, and 28 weeks

- **Number of abortions** & **Maximum week of abortion**: These features record how many abortions a participant experienced, and the maximum week in which an abortion was conducted.

  They were created by extracting information from the binary pregnancy loss column, and its extension, pregnancy loss text, in which women described previous pregnancy losses.

- **Number of miscarriages** : Records the number of miscarriages a participant previously had.

  As the above features, the numerical value was extracted from the text description of the Recent miscarriage column.

- **Problems previous pregnancy** : Binary feature recording whether a woman had problems during previous births.

- **Number of premature weeks in previous births**: This feature records the number of weeks early a previous child was born if that child was born prematurely.

- **Gestational weight gain features** : 3 features recording the weight gain from 12 weeks to 20 weeks, from 20 weeks to 28 weeks, and the total weight gain during the pregnancy

This step involved creating functions to extract numerical inputs from text data (an example shown in Algorithm 1)

Algorithm 1: Pseudocode for extracting numbers

```
Function extract_numbers(text)
    numbers ← findall('(?:0|[1-9][0-9]?|20)',
      text);
    if numbers is not empty then
        return numbers[0];
    else
        return 1;
```

4.6 Demographic Data Overview

After removing all demographic features with more than 50% missing values , the remaining relevant features were :

- Breadwinner : categorical feature denoting the breadwinner in the family

- Education: categorical feature denoting the education level of the participant at baseline

- Maritalstatus : categorical feature denoting the marital status of the participant
- Samepartner: dichotomous feature which indicates whether the participant is in the same relationship with the same partner as the previous pregnancy
- Age: numerical feature

4.7 Overweight Classification Methodology

Since the focus of the analysis is to determine risk factors leading to overweight/obese children, overweight was deemed as the target variable. To classify children within the dataset as overweight/obese or non-overweight/obese, a careful analysis must be conducted to uncover the underlying trends within the data. The follow-up to the pregnancy data includes measurements of length and weight for children at the ages of six months, one year and six months, 2 years, 2 years and six months, 3 years, 3 years and six months, 4 years, and 4 years and six months. A participant's child is classified as overweight if they are overweight compared to their peers at any age. For instance, if a child became overweight at 3 years old but was no longer overweight at 4 years old, they would still be classified as overweight. The next step involves determining at which age a child can be classified as overweight. Below, two ways of classification will be presented for children above 2 years old and children below 2 years old.

4.7.1 Ages at which children had their highest BMI

As described above, children were classified as overweight whether, at some point in time, they were deemed overweight. Thus, it can be interesting to take a look at the ages at which children had their highest BMI measurement.

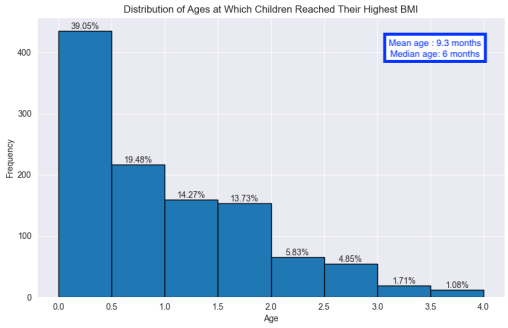

Figure 3: Distribution of ages at which children had their highest BMI

As shown in 3, around 86% of the children had their highest BMI when they were younger than 2 years old, indicating that the follow-up data is limited.

4.7.2 Overweight classification for children **older than 2 years old**

Classification for children older than 2 years old was done according to the classification provided within the Brabant Study, [1](shown in Table 4).

Table 4: BMI cut-off points (kg/m²) for overweight and obesity for boys and girls [17]

Age (years)	Boys Overweight (kg/m²)	Boys Obesity (kg/m²)	Girls Overweight (kg/m²)	Girls Obesity (kg/m²)
2	18.4	20.1	18.0	19.8
3	17.9	19.6	17.6	19.4
4	17.6	19.3	17.3	19.2
5	17.4	19.3	17.2	19.2

4.7.3 Overweight classification for children **younger than 2 years old**

Children under 2 years old couldn't be classified as overweight or not using established cut-off values: a method of classification based on the data needed to be found. Thus, the next step implied analyzing whether the sample data is representative of the Dutch population. It was done by comparing trends for overweight and obesity referring to data shown in 4 and trends within the sample data.

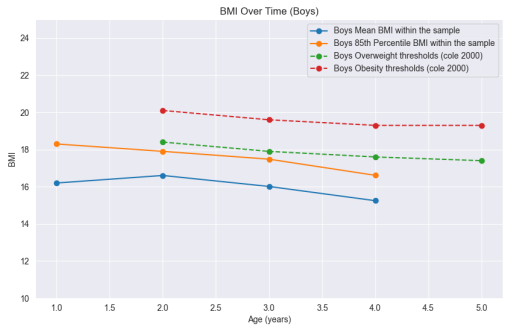

Figure 4: Trends for Population and Sample Values: Mean BMI, 85th percentile, Overweight, and Obesity Cutoff Values Among Boys

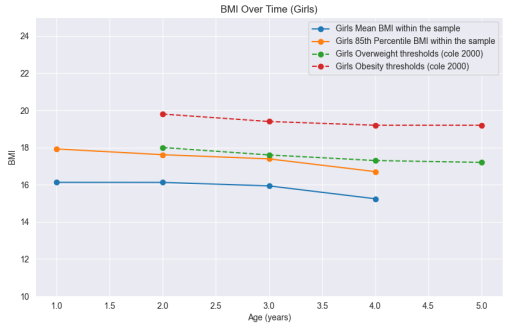

Figure 5: Trends for Population and Sample Values: Mean BMI, 85th percentile , Overweight, and Obesity Cutoff Values Among Girls

As shown in the above plots, the trends for Dutch BMI cutoff population values for overweight and obesity (the ones displayed in 4) are similar to the trends for mean BMI values within the dataset. Moreover, the cutoff values in Table 4 exhibit trends that are comparable to the trends for the 85th percentile values within the sample.

Therefore, it can be assumed that the Brabant Study sample data shares similarities with the population.

Thus, overweight classification for children under 2 years old is based on the sample data. A child is classified as overweight if their BMI is among the top 15% for their age group in the sample. This classification method differs from the one provided by the World Health Organization, but it was chosen due to the nature of the data. It was considered more appropriate to assess the data with the help of benchmarks used by the Brabant Study as well [1], since the two analysis are both conducted within the Netherlands.

4.8 Final Data Cleaning and imputation

Finally, all 3 sets of features mentioned above were merged on Participant Number, which is the unique identifier of the dataset. Participants with missing values in the target variable(overweight classification) were eliminated. Moreover, outliers such as strange BMI values were removed(when measures that compose the BMI, such as weight or length were among the outliers). Thus the number of participants was reduced in the data cleaning step by 214(from 1328 to 1114). The next issue is how missing data should be dealt with.

4.8.1 Mean-Mode Imputation

This method involves feeling missing numerical data with the mean for that feature, and the missing categorical values with the mode of the categorical feature.

4.8.2 MICE imputation

The second method proposed for filling missing values is employing the MICE method in R, as it was done in [16] for numerical and categorical variables.

5 Prediction Analysis

This method aims to properly classify a child as overweight or not based on his mother's pregnancy data and mental health. The focus was shifted from investigating obesity to investigating overweight because of the data limitations. While for obesity there are 100 obese children out of 1300, taking overweight as the target variable would increase the positive class to around 300 overweight children out of 1300 children. With a test size of 15%, the train test split would look as in Figure 8

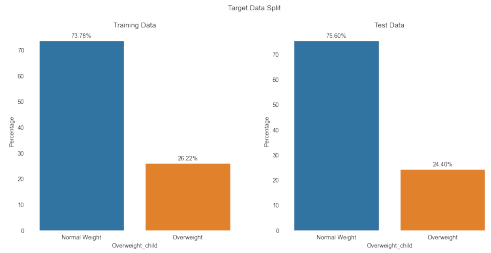

Figure 6: Train-Test split for target Data with test size of 0.15

Preliminary predictive analysis included employing two models: a Logistic Regression Model and an XGboost Model. For the models below, missing data was filled using the mean-mode imputation method. Furthermore, categorical data was one-hot encoded, while numerical data was scaled, using a standard scaler.

The prediction models will have as independent variables the following features:

- **Categorical Features** : Type_D_personality, Breadwinner_12, Education_12, Maritalstatus_12, Plannedpreg_12, Locationdelivery_12, Problemspreg_12, Painmanagement_12, Supplements_12, Medication_12, Medication_unreg_12, Problemspreg_20, Doctorvisit_20, Medication_20, Supplements_20, Problemspreg_28, Doctorvisit_28, Supplements_28, Medication_28, Chronicdisease_TOT_12, BMI_CAT_12, BMI_CAT_20, BMI_CAT_28, mental_health_issue, mental_health_treatment, Guided_by_obstetrician, Problems_prev_preg

- **Numerical Features**: GWG_total, BMI_28, Social_media_use_28, Depression_20, Pregnancy_distress_12, GWG_2ndperiod, Difficulties_emotion_regulation, Social_media_use_12, Anxiety_28, Referred_gynaecologist_at_n_weeks, Smoking_per_day_week_12, Bonding_pre_birth, FERRH_28, FT4H_12, nr_of_weeks_premature, HCGBH_12, Morning_sickness_20, Depression_12, Psychological_complaints_8wPP, Work_mental_health_12, Work_performance_12, BMI_highest_child, IL6H_28, Age_depression, PLGFH_20, Relationship_satisfaction_20, max_week_abortion, PLGFH_12, Interaction_newborn_baby, ATPOH_20, IL6H_12, Weightpreg_20, BMI_20, Nausea_up_to_n_weeks, Smoking_per_day_week_20, Age_highest_BMI_child, Partner_involvement_28, Age_anxiety, Relationship_satisfaction_28, ATPOH_12, Personality_BFI, Bonding_post_birth, Depression_28, HCGBH_20, Anxiety_12, Mindfulness, BMI_12, Partner_involvement_20, Weightpreg_28, ATPOH_28, PLGFH_28, Work_performance_28, FT4H_20, Anxiety_20, IL6H_20, FERRH_12, GWG_1stperiod, TSHH_28, Pregnancy_distress_20, Social_media_use_20, nr_abortions, Medication_unreg_28, HCGBH_28, FERRH_20, Smoking_per_day_week_28, Partner_involvement_12, Social-support, Age_12, FT4H_28, TSHH_12, Pregnancy_distress_28, Morning_sickness_12, Morning_sickness_28, Relationship_satisfaction_12, Work_mental_health_28, TSHH_20

5.1 Importance of False Negatives and False Positives in Overweight Prediction

Accurately predicting overweight status in children is essential for implementing early interventions as well as preventing future health issues. In this context, it is very important to understand the impact of false negatives and false positives

False negatives occur when a child who is overweight is incorrectly classified as not overweight. The consequences of false negatives could be missed interventions: parents with children at risk of becoming overweight remain unaware, thus not adjusting the child's diet accordingly.

False positives occur when a child who doesn't face any risk of becoming overweight is incorrectly classified as overweight. In this context, false positives are not detrimental, as a healthy diet is always beneficial: parents might take action assuming their child is at risk, leading to better nutrition. One of the only potential downsides is the stress a child might develop from having to follow a stricter diet, but this is relatively insignificant.

In conclusion, while both false negatives and false positives in overweight prediction have consequences, false negatives are more detrimental than false positives and, thus less desired.

5.2 Logistic Regression Model Evaluation

The model achieves a good overall accuracy, but it has some limitations. Due to the unbalanced data split between the two

classes, the non-overweight class being over-represented(that is the nature of the data), a lot of emphasis must be put into classifying over-weight. In the situation of informing overweight, False Negatives(falsely classifying a child as having no risk for being overweight) could be considered more harmful than False Positives. Thus, the model's main drawback is having more False Negatives than True Positives as well as having more False Negatives than False Positives.

Table 5: Classification Report Logistic Regression

	Precision	Recall	F1-Score	Support
0	0.83	0.89	0.86	127
1	0.56	0.44	0.49	41
Accuracy		0.78		168
Macro Avg	0.70	0.66	0.68	168
Weighted Avg	0.77	0.78	0.77	168

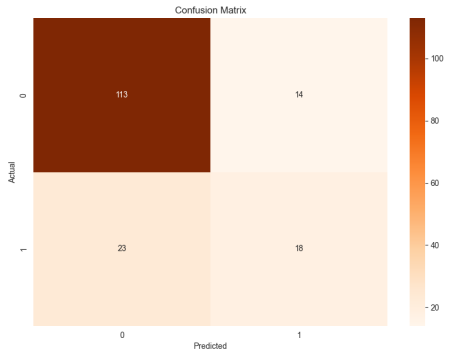

Figure 7: Confusion Matrix Logistic Regression

5.3 XGboost Model Evaluation

The XGboost model outperforms logistic regression as highlighted in 6: it achieves better performance in every kind of comparable metric, with an impressive 0.92 weighted f1 score.

Nevertheless, a major improvement is achieving more True Positives than False Negatives, thus better classifying overweight children as overweight. Unfortunately, it suffers from the same issue as logic regression , with more False Negatives than False Positives, which is not desired in the situation of informing overweight(8).

Table 6: Classification Report XGboost

	Precision	Recall	F1-Score	Support
0	0.92	0.99	0.95	127
1	0.97	0.73	0.83	41
Accuracy		0.93		168
Macro Avg	0.94	0.86	0.89	168
Weighted Avg	0.93	0.93	0.92	168

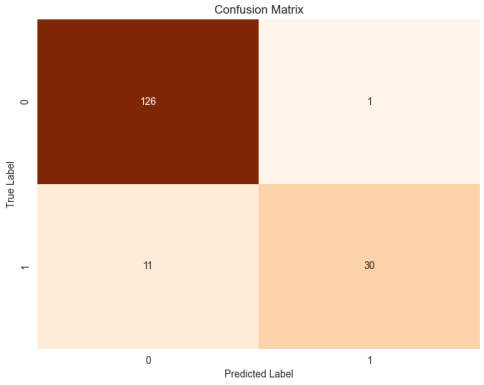

Figure 8: Confusion Matrix XGBoost

6 Causal Analysis

This analysis seeks to identify potential risk factors for overweight in preschool-aged children. Causal analysis can be thought of as an adequate tool to identify determinants of overweight.

Compared to predictive analysis, causal analysis lays out a more in-depth understanding of the risk factors for overweight, identifying specific factors that might influence a child becoming overweight. While predictive modeling can be a good screening tool for estimating the likelihood of being overweight, causal analysis is more beneficial for prevention efforts. Previous research like [6] also covered medical factors leading to obesity in children.

6.1 Logistic Regression

To identify significant factors for overweight in pre-school-aged children, a logistic regression was run, with the dependent variable as overweight and the independent variables, all the other features in the dataset. For categorical features with more than 2 unique values(not only 0 and 1- binary), dummy variables were created. Below is an overview of the significant factors within the model(with `p-value < 0.05`). An overview of the significant predictors found can be seen in Table 7

Significant Predictors	
Variable Name	**Measures**
Depression_20	depression level at 20 weeks of pregnancy
FERRH_20	indicates stored iron levels and inflammation
Problems_prev_preg	binary variables assessing if a woman had issues during the previous pregnancy
Age_highest_BMI_child	age at which a child had the highest BMI
Locationdelivery_12_1	binary - mother had a home-birth
Locationdelivery_12_2	binary - mother had a hospital birth
Painmanagement_12_1	binary - mother used pain killers during pregnancy

Table 7: Significant predictors identified in the logistic regression.

6.2 IV Regression

Instrumental variable (IV) regression is a tool used to estimate causal relationships. A key advantage of IV regression over logistic regression for causal analysis is effectively addressing endogeneity issues, that can lead to biased and inconsistent regression. This is done by implementing a two-stage regression.

6.2.1 First Stage Regression

The first stage aims to isolate variation in the endogenous variables, attributable to the instruments.

In this implementation, the endogenous variables are the significant features found in the logistic regression, while the instruments are the set of all the remaining features in the dataset.

The fitted values of this first stage capture the variation in the endogenous variables, explained by these instruments. They are then used as independent variables in the second stage of regression.

6.2.2 Second Stage Regression

In the second stage, the dependent variable (overweight classification for children) is regressed on the predicted values mentioned, instead of the original endogenous variable. By using the predicted values from the first stage, the endogenous problem is solved because these values are uncorrelated with the error term. Thus, the obtained coefficients can be deemed unbiased and consistent.

6.2.3 Results

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.04527	0.06828	0.663	0.5075
Depression_20	-0.02895	0.12266	-0.236	0.8135
FERRH_20	-0.19425	0.23058	-0.842	0.3997
Problems_prev_preg	0.16132	0.07627	2.115	0.0346 *
Locationdelivery_12_1	-0.01718	0.11000	-0.156	0.8760
Locationdelivery_12_2	-0.10661	0.05458	-1.953	0.0510 .
Age_highest_BMI_child	1.38764	0.17607	7.881	7.72e-15 ***
Painmanagement_12_1	0.10806	0.04188	2.580	0.0100 *

Table 8: IV Regression coefficients with standard errors, t-values, and p-values. Significance codes: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$), . ($p < 0.1$)

Significant causal relationships between overweight and the other features

- Problems in Previous Pregnancies: The variable $Problems - prev - preg$ has a significant positive coefficient, indicating that mothers who had issues during previous pregnancies are at a higher risk of having an overweight child.
- Pain Management: The variable $Painmanagement - 12 - 1$ is also significant, suggesting that the use of pain management pills during the earlier stages of pregnancy is associated with an increased risk of having an overweight child.
- Age-highest-BMI: The variable $Age - highest - BMI - child$ is also significant. Thus, it is suggested that the age at which a child has his highest BMI positively affects the likelihood of him being overweight

7 Limitations

7.1 Generalizability

7.1.1 Sample size and representation

Although the sample size is sufficient to draw pertinent conclusions, the main issue within the data is the underrepresentation of marginalized groups, particularly the very low percentage of uneducated individuals. This thus leads to a significant misrepresentation of an entire population segment.

As demonstrated in Devaux et al. (2011) [18], a higher level of education is associated with a reduced likelihood of being obese. Furthermore, Singh et al. (2014) [19] highlighted that a mother's education level significantly influences her children's BMI. Thus, education can be considered a critical variable in obesity research.

However, within this study, the participants analyzed predominantly have high levels of education, as illustrated in Figure 9. In the sample, the percentage of women with their highest form of education as secondary/high school is 3.92%, while the number of women having as their highest level of education University is 67.55%.

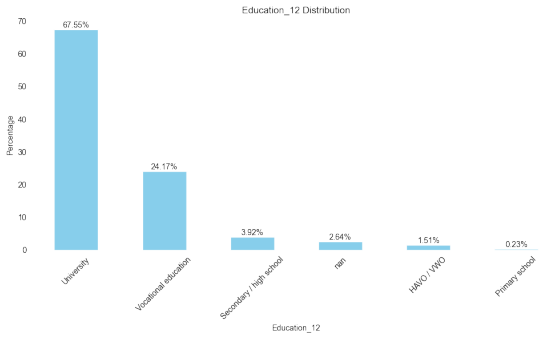

Figure 9: Mother Education Level Distribution

This lack of diversity in education levels within the collected data limits the comprehensiveness of all analyses. The over-representation of highly educated mothers restricts the ability of this study to generalize findings.

Thus, future research should aim to include in the data collection phase a more diverse sample of women, regarding education levels.

7.1.2 Geographic and Cultural Specificity

The Brabant study was conducted within a specific region: South-East Brabant, Netherlands [1]. Given that this paper is based on data collected within the Brabant Study, the results may not be applicable for other geographic regions, with different socio-economic and cultural backgrounds.

7.2 Methodological Constraints: Feature Engineering

While feature engineering was utilized to combine related features, create new variables, and address missing values to uncover broader trends in the data, this approach also has some drawbacks. One major drawback is the risk of introducing bias through feature aggregation.

8 Conclusion

This paper aims to investigate the research questions formulated in subsection 1.8, highlighting the limitations of the analysis as well.

8.1 Research question 1: How does predictive modeling work in identifying the risk of childhood overweight

As outlined in the predictive modeling section, the XGBoost model outperforms simple logistic regression in classifying children as overweight. Implementing such a model requires a careful breakdown of its use and implications.

Firstly, discussing the relative harms of false negatives versus false positives in the current use case is important. In this scenario, false negatives are more harmful than false positives. This is because preventing obesity can be beneficial regardless of whether a child is misclassified as at risk for becoming overweight. A false negative, however, could result in a parent not understanding the predisposition of his child to being overweight, leading to a lack of necessary intervention.

Table 9: Performance Metrics of the XGBoost Model

Metric	Value (%)
True Positive Rate (TPR)	73.17
True Negative Rate (TNR)	99.21
False Positive Rate (FPR)	0.79
False Negative Rate (FNR)	27.00

The XGBoost model exhibits a high true positive rate (73.17%) and an even higher true negative rate (99.21%), underlining its strong performance in correctly identifying both positive and negative cases. However, it is essential to acknowledge that the false negative rate stands at 27%, as illustrated in Table 9.

In the context of developing a recommendation system to alert mothers about the risk of their child becoming overweight, the high false negative rate is particularly concerning. A false negative might cause a parent to underestimate her child's risk, resulting in insufficient attention to the child's nutritional needs. To mitigate this, it is necessary to implement the recommendation system with the support of nutritionists who can provide expert guidance and explain the model's limitations.

8.2 Research Question 2: What causal relationships exist between childhood overweight and maternal characteristics?

In addition to maternal characteristics such as maternal BMI, gestational weight gain, and maternal pre-pregnancy BMI, which have been identified as risk factors in previous research [13], our analysis has revealed additional significant factors that contribute to childhood overweight.

Specifically, the use of pain management pills by mothers during pregnancy and the presence of complications in previous pregnancies have emerged as notable predictors of childhood overweight. The $Problems - prev - preg$ variable encompasses all sorts of issues a woman might have during previous pregnancies, thus a breakdown of specific issues is not possible within the current framework.

To conclude, the most important causal relationship found within the dataset is between a mother having issues during previous pregnancies(problems previous pregnancies) and her child's likelihood of becoming overweight . Unfortunately , the data doesn't allow for further analysis into the exact complications, due to the high number of missing values. Thus, future research needs to put an emphasis on examining this relationship and identifying the specific issues involved.

REFERENCES

1. Meems, M. *et al.* The Brabant study: design of a large prospective perinatal cohort study among pregnant women investigating obstetric outcome from a biopsychosocial perspective. *BMJ Open* **10,** e038891. https://bmjopen.bmj.com/content/10/10/e038891 (2020).
2. World Health Organization. *Obesity and overweight* Accessed: 2024-04-03. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (2024).
3. National Institute of Diabetes and Digestive and Kidney Diseases. *Health Risks of Overweight and Obesity* https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/health-risks. Accessed: 2024-05-23. 2024.
4. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *New England Journal of Medicine* **377,** 13–27. https://www.nejm.org/doi/full/10.1056/NEJMoa1614362 (2017).
5. De Deus Mendonça, A. *et al.* Ultraprocessed food consumption and risk of overweight and obesity: the University of Navarra Follow-Up (SUN) cohort study. *The American Journal of Clinical Nutrition* **104,** 1433–1440. ISSN: 0002-9165. https://www.sciencedirect.com/science/article/pii/S0002916522046767 (2016).
6. Pinhas-Hamiel, O. *et al.* The Global Spread of Severe Obesity in Toddlers, Children, and Adolescents: A Systematic Review and Meta-Analysis. *Obes Facts* **15.** Epub 2022 Jan 11, 118–134. https://doi.org/10.1159/000521913 (2022).
7. Duncan, D. T. *et al.* Characteristics of Walkable Built Environments and BMI z-Scores in Children: Evidence from a Large Electronic Health Record Database. *Environmental Health Perspectives* **122,** 1359–1365. https://doi.org/10.1289/ehp.1307704 (2014).
8. Park, M. H., Falconer, C., Viner, R. M. & Kinra, S. The impact of childhood obesity on morbidity and mortality in adulthood: a systematic review. *Obesity Reviews* **13.** First published: 26 June 2012, Citations: 500, 985–1000. https://doi.org/10.1111/j.1467-789X.2012.01015.x (2012).
9. Vieweg, V. R. *et al.* Correlation between high risk obesity groups and low socioeconomic status in school children. *Southern Medical Journal* **100.** Accessed 15 June 2024, 8+. https://link.gale.com/apps/doc/A158957732/AONE?u=anon~a207ad1&sid=googleScholar&xid=e79fd8cc (2007).
10. European Commission. *Council Recommendation Establishing a European Child Guarantee* Accessed: 2024-06-15. https://ec.europa.eu/social/main.jsp?catId=1428&langId=en#JAF.
11. Government of the Netherlands. *National Plan: Dutch Situation Regarding Policy on Child Poverty* Accessed: 2024-06-15. https://ec.europa.eu/social/BlobServlet?docId=25518&langId=en.
12. Wang, Y., Min, J., Khuri, J. & Li, M. A Systematic Examination of the Association between Parental and Child Obesity across Countries. *Systems-Oriented Global Childhood Obesity Intervention Program, Fisher Institute of Health and Well-Being, and Department of Nutrition and Health Sciences, College of Health, Ball State University, Muncie, IN; Systems-Oriented Global Childhood Obesity Intervention Program, Department of Epidemiology and Environmental Health, School of Public Health and Health Professions, University at Buffalo, The State University of New York, Buffalo, NY; and University of Redlands, Redlands, CA.* Available online 5 May 2017, Version of Record 11 January 2023. https://www.sciencedirect.com/science/article/pii/S2161831322006755 (2017).
13. Wang, Q. *et al.* Explorations on risk profiles for overweight and obesity in 9501 preschool-aged children. *Obesity Research & Clinical Practice* **16,** 106–114. https://doi.org/10.1016/j.orcp.2021.10.010 (Mar. 2022).
14. Liang, J. *et al.* Association between both maternal prepregnancy body mass index/gestational weight gain and overweight/obese children at preschool stage. Chinese. *Zhonghua Liu Xing Bing Xue Za Zhi* **40,** 976–981. https://pubmed.ncbi.nlm.nih.gov/31484264/ (2019).
15. Lim, H., Lee, H. & Kim, J. A prediction model for childhood obesity risk using the machine learning method: a panel study on Korean children. *Sci Rep* **13,** 10122. https://doi.org/10.1038/s41598-023-37171-4 (2023).
16. Stoitsas, K. *et al.* Clustering of trauma patients based on longitudinal data and the application of machine learning to predict recovery. *Sci Rep* **12,** 16990. https://doi.org/10.1038/s41598-022-21390-2 (2022).
17. Cole, T., Bellizzi, M., Flegal, K. & Dietz, W. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ* **320,** 1240–1243. https://pubmed.ncbi.nlm.nih.gov/10797032/ (May 2000).
18. Devaux, M. *et al.* Exploring the Relationship Between Education and Obesity. *OECD Journal: Economic Studies* **2011.** https://read.oecd-ilibrary.org/economics/exploring-the-relationship-between-education-and-obesity_eco_studies-2011-5kg5825v1k23#page1 (2011).
19. Singh, D., Goli, S. & Parsuraman, S. Association between obstetric complications & previous pregnancy outcomes with current pregnancy outcomes in Uttar Pradesh, India. *Indian Journal of Medical Research* **139,** 83–90. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3994745/ (Jan. 2014).