Zsófia Práger

Longitudinal analyses of childhood obesity factors

Master's Thesis

Supervised by Prof. Beatriz López and Dr. Judit Bassols

Masters's Degree in Biomedical Data Science

Budapest, 2024

Prof. Beatriz López certifies that the student Zsofia Prager has elaborated the work under her direction and she authorizes the presentation of this Master's Thesis for its evaluation.

Advisor signature: .....

Abstract

This master thesis investigates factors contributing to childhood obesity through a longitudinal analysis of data from hospitals in Catalonia, building on previously conducted experiments with this dataset. The utilized dataset consisted of a cohort of pregnant women and their infants. The children were followed up until the age of 5 detecting their childhood obesity status. The collected variables consisted of parental information, including variables related to the father, the mother pre-, and during the pregnancy alongside with information about the infant's first year of life.

The aim of the study was to create a predictive model, that would be able to correctly classify the obesity status of children and identify the key variables responsible for determining the differentiation. Two new experimental factors were introduced: transformation of raw data along time into increments, and representation of data with sequential patterns. To that end, a sequential pattern mining algorithm has been applied for learning sequential patterns. A total of 150 distinct patterns related to obesity status at age five were identified and analyzed using three classification algorithms: Logistic Regression, Decision Tree, and Random Forest. The most explanatory patterns were examined to assess the relevance of the variables.

The sequential patterns found relevant in this study were additionally tested and the results were compared against previous findings. This comparison demonstrated the weakness of these new methods in the context of childhood obesity research, regarding other state-of-the art approaches.

Keywords: Childhood obesity detection, Artificial Intelligence, Sequential Pattern Mining, Logistic Regression, Decision Tree, Random Forest

Contents

List of Figures

List of Tables

1 Introduction

In this chapter the main motivations for this work are described, alongside with it's main contributions.

1.1 Motivation

As described in [1], a research conducted at 2022 by the World Health Organization (WHO) revealed that 37 million children under the age of 5 were overweight. In Europe, this translates to every third child being overweight, according to WHO's European Childhood Obesity Surveillance Initiative.

Based on the work of [2], childhood obesity imposes many later-life problems for individuals, including both physiological and psychological issues. These serious illnesses include Type 2 diabetes, high blood pressure, stroke, mental health disorders, and many others. [3]

These findings and scientifically proven consequences highlight the critical importance of investigating the factors contributing to the development of childhood obesity to address this growing public health concern effectively.

Dr. Judit Bassols Casadevall leads a research group focused on obesity-related questions at the Institut d'Investigació Biomèdica de Girona (IDIBGI). This work was developed in close collaboration with Dr. Bassols and the research center, with data collection and provision facilitated by them. The collected data was a cohort of pregnant women and their children. The data consisted of information about the childrens' father, mother (before and during pregnancy) and the children themselves. The final dependent variable was concerned with the last observation at the children' 5 years of age obesity status.

The main motivation for this research has been aligned with addressing these problems. It is undoubtedly essential to explore new methods and solutions to effect positive change. This internal motivation, combined with the data science and machine learning knowledge acquired during the writer's master's studies, has driven the commitment to this project. By leveraging advanced machine learning and sequential data mining, the aim is to uncover actionable insights that can contribute to the prevention and reduction of childhood obesity.

1.2 Objectives and Hypotheses

This given research focuses on identifying factors contributing to obesity in 5-year-old children using artificial intelligence and my aim is to develop methods that can be utilized in future scientific studies. To achieve this, novelty methods in data pre-processing, sequential pattern mining and various machine learning algorithms

were employed to uncover and understand patterns in the data.

The primary objective of this work was to develop a machine learning-based solution for the early detection of children at high risk of obesity. This involved introducing effective data manipulation techniques, pattern mining and creating predictive models to identify such individuals at a young age.

Sequential pattern mining was incorporated to leverage the longitudinal nature of the collected data. Previous studies conducted with the utilized data overlooked the time dimension. This work analyzed data from sequential visits, and aimed to keep the longitudinal manner of the data which was crucial to revealing underlying patterns.

The hypothesis was that combining sequential pattern mining with other machine learning algorithms can effectively identify factors contributing to obesity in 5-year-old children. It was expected that these identified factors align with findings from clinical studies. It was anticipated that employing sequential pattern mining and initial data-preprocessing the obtained results outperform findings from previous studies.

1.3 Thesis Structure

Chapter 2 describes the already available literature regarding the current research on childhood obesity factors, the methods employed for the prediction and identification of obesity with machine learning, the current sequential pattern mining research and summarizes the previous work that has been done to the used database.

Following this, the key concepts and theory regarded to sequential pattern mining and the used machine learning models are defined and described in Chapter 3.

In Chapter 4, the performed data analysis is described. The methodology followed during the research is shown while presenting the actual dataset, data preprocessing and exploratory data analysis. The steps of the sequential pattern mining and the machine learning modelling and the methods of how the results were extracted are also provided.

In Chapter 5 the results are presented, and in Chapter 6, the discussion of the work is provided, with regards of the initial hypothesis. Limitations and future work is described in Chapter 7, to propose other ways to obtain better results in the future. Chapter 8 sumarizes the work and the results. The last part of the work, Chapter 9 describes the ethical-social impact, the sustainability and the diversity questions related to the proposed research.

2 Literature Review

In this chapter, currently available research is described on the factors leading to childhood obesity. Additionally, a summary of previous approaches is included on detecting obesity and childhood obesity, as well as sequential pattern mining algorithms used in similar problems.

2.1 Current Research on Childhood Obesity Factors

It is unanimously accepted in the scientific community that the main cause of obesity is the imbalance between energy intake and expenditure [1]. Although, genetic background and several prenatal, postnatal, and early life factors are all considered risk factors for childhood obesity alongside with various lifestyle choices.

A cohort study conducted in the UK identified eight putative risk factors associated with later life obesity [4]. This study provided evidence that both early life and prenatal factors, along with lifestyle factors are significant. The identified factors were parental obesity, early BMI of infants (by 43 months), high hours of television watching, the prevalence of catch-up growth in children, standard deviation for weight at 8 months and 18 months, weight gain in the first year, birth weight, and short sleep duration.

A systematic review of 14 observational studies also reported the association between maternal smoking and childhood overweight [5]. Based on 84.563 distinct children an elevated risk for overweight was observed for individuals at age of 3-33 with smoking mothers.

In [6], a multivariate regression analysis showed that high gestational weight gain alongside with gestational obesity is associated with higher child body mass index. Women who show adequate or excessive weight gain during pregnancy have children with higher risks of obesity. This finding was confirmed again in a population-based cohort study that showed a positive association between pregnancy weight gain, high birth weight and the likeliness of obesity [7].

Multivariate regression analysis showed significant association between maternal gestational hypertension and the later obesity status of children [8]. The given study observed adolescents aged 12-25.

Another systematic overview study found that babies whom show high body mass index or accelerated weight gain during infancy are at increased risk of later life obesity [9].

In [10], postnatal parental smoking status and the association between obesity was researched in a cohort study from Ireland. The study categorized the participants in four groups based on the exposure status of smoking: neither of the caregivers smoked, only primary caregiver smoked, both caregivers smoked. The

study concluded that the expenditure of either smoking of primary caregivers or both parents is significantly related to obesity at age 3 and age 5.

Prenatal smoking is also associated with childhood obesity. The study [11], concludes that maternal smoking leads to obesity independent to maternal pre-pregnancy BMI and the genetic predisposition of extreme obesity.

Two meta-analyis from [12] and from [13], comparing breastfeeding to bottle feeding found that breastfeeding was linked to a consistent protective effect against obesity. However this is still a controversial area of research, due to many possibly prevalent confounding factors and other influencing agents such as socio-economic status or culture. [14], also found a dose-dependent negative association between duration of breastfeeding and the risk of obesity.

In summary, multitude of distinct factors have been clinically and statistically proven to have influence on the occurrence and development of childhood obesity. These influencing factors include parental lifestyle choices before and during pregnancy, birth related factors and early life developments aspects.

2.2 Obesity Prediction Based on Machine Learning Techniques

The use of machine learning models in detecting obesity has been widely explored in previous research, incorporating various data choices and model selection methods. A systematic review paper [15], has shown that the number of publications and citations combining keywords of obesity, overweight, and machine learning has been increasing over the last ten years up until 2021. The vast majority, around 70%, of these publications were found to be published in the 2019-2021 period, by the date of the publication of the study (2021), indicating a growing scientific focus and interest in the topic.

Another scoping review demonstrated that diverse machine learning and deep learning models are utilized for obesity detection [16]. The majority of the reviewed publications indicated that AI models achieve higher prediction accuracy compared to traditional statistical methods. The study also highlighted a growing trend in the inclusion of deep learning methods in obesity research.

A prediction model proposed in [17], aimed to provide a computer aided diagnosis (CAD) solution for hospitals and medical professionals to predict the obesity status of patients. Out of the four tested models, the gradient boosting algorithm outperformed the other approaches, resulting in an accuracy score of 99.05%.

Machine learning has been widely used and researched specifically on childhood obesity predictions. In [18], seven different algorithms were employed to predict obesity in children aged 2 to 7. The data used was based on the children's electronic

health records up to age 2. This study included 27,203 patients, making it the largest cohort used for such a study at the time. The results indicated that the XGBoost algorithm produced significantly superior results in obesity detection.

A similar study was conducted using Israeli electronic health records, primarily focusing on changes in children's BMI over time [19]. The cohort selection criteria were based on having an adequate number of medical checkups where weight and length measurements were taken, as well as the child's obesity status at 5 years old. Once again, the XGBoost algorithm was used for obesity prediction. The method demonstrated promise in identifying and highlighting the importance of BMI development in the first 2 years of life in relation to later obesity.

2.3 Sequential Pattern Mining Solutions for Similar Research Questions

Sequential pattern mining (SPM) algorithms have been utilized in longitudinal studies for predicting diagnoses for various different diseases. Demonstrating the capabilities of such predictions imply that using SPM for obesity recognition could be a potentially effective approach.

One interesting work is [20], that introduced temporal condition pattern mining for sparse, coded EHR data. The study applied pattern mining on pediatric asthma diagnoses. The work prevailed that using such methods might help the discovery of previously unknown associations between various factors and the development of asthma.

In [21], three different pattern mining algorithms were tested and applied on Chinese EHR data. The aim of the study was to detect treatment and medication categories used for children with pneumonia. The algorithms were able to identify five distinct medication use patterns and two treatment patterns .

In [22], the SPADE algorithm was utilized for obesity recognition based on EHR data too. The method utilized previous diagnoses that were present in at least 1% of the observed cases in the target population. Two conditions were identified as significantly more common at children diagnosed with obesity, these were asthma and allergic rhinitis.

These examples of pattern mining in healthcare justify the relevance and show the promise the method provides. It is able to detect underlying patterns, that have previously been overlooked even by professionals, and enable researchers to focus on new research directions and topics.

2.4 Previous Work on the Data

This work builds upon previous research conducted at the Institut d'Investigació Biomèdica de Girona (IdiBGi). Several papers and master's theses have been written based on the data used and the methods employed in this current study.

In [23] and [24], initial data imputation and basic data manipulation were combined with machine learning methods applied to the entire initial dataset. Additionally, [23] introduced itemset mining for the detection of frequent patterns. This work has proven that incorporating itemset mining is a viable approach for modelling obesity with longitudinal data, as using the mined patterns showed better results than using the absolute values for the modelling.

Although this work is similar to these previous experiments and studies, sequential pattern mining, new modeling approaches and data manipulation methods were introduced to further advance the research. This includes additional techniques for data preparation (creation of the increments instead of absolute values) aiming to enhance the accuracy and effectiveness of the models in identifying factors contributing to obesity.

3 Background

The aim of this chapter is to present the theoretical background of the used methods. The description of the classification problem and the different algorithms is based on the book of Bonaccorso [25]. The overview includes, classification in machine learning and the models that were included in this work. The chapter also describes a high level introduction of sequential pattern mining and the used algorithm VEPRECO,

3.1 Classification Models and Machine Learning Techniques

Classification in machine learning is a fundamental and common task. It arises when the dependent variable is categorical. As a supervised learning task, classification relies on labeled data during training, meaning the expected labels are provided to train the model.

Different classification metrics can be used for assessing the performance of the model. The used criterion are model agnostic metrics, including accuracy, precision, recall and F1 scores.

The accuracy score identifies the number of correctly classified items out of the total number of samples. It is often represented and interpreted in percentages.

The remaining 3 metrics are mostly aimed at identifing where the miss-classification happen. They are all closely related to the confusion matrix, which includes the number of correctly and incorrectly classified negative and positive samples.

The proportion of true positives out of all values classified as positive is the precision metric. It captures the positiveness of the classification. Recall stands for the rate of true positives out of all the originally positive values(true positives, false negatives). The F1 score is the harmonic mean between precision and recall.

There are numerous machine learning algorithms to build classification models, each based on different principles. The algorithms used in this work include logistic regression (LR), support vector machines (SVM), induction of decision trees (DT), random forest (RF), and extreme gradient boosting (XGBoost). The performance of all of them depend on several parameters that are identified by using the grid search algorithm.

3.1.1 Logistic Regression

LR is a widely used method in applied statistics and data analysis. It was proposed in the late 1960s making it an older but reliable approach in classification. A key aspect of logistic regression is that the outcome is dichotomous, it can take up only 2 possible values.

The main underlying method in logistic regression is based on the calculation of

the probability of the sample to belong to the classes. For calculating this probability a threshold function, the sigmoid function is needed. This function can be given as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

The resulting plot of the function shows a linear line with curved ends, resembling an S-shaped curve with values between 0 and 1. These values could be interpreted as probabilities.

Logistic regression applies a logit transformation on the dependent variable. This logit is modeled as a linear combination of the predictor variables. The model equation is the following:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where:

- p is the probability of the dependent variable being 1 (e.g., the event occurring).
- $1 - p$ is the probability of the dependent variable being 0 (e.g., the event not occurring).
- β_0 is the intercept term, which represents the log-odds of the dependent variable being 1 when all independent variables are 0.
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients that measure the change in the log-odds of the dependent variable for a one-unit change in the corresponding independent variables $X_1, X_2, \ldots, X_k$, respectively.
- $X_1, X_2, \ldots, X_k$ are the independent variables (also called predictors or features).

Similarly to ordinal least squares (OLS), logistic regression has assumptions about the data, which are the following:

- **Assumption 1:** Binary dependent variable.
- **Assumption 2:** Independence of the observations from one another.
- **Assumption 3:** No multicollinearity between predictors.
- **Assumption 4:** Linearity of the logit dependent variable and the predictor variables.
- **Assumption 4+1:** Large sample size.

LR has many advantages and disadvantages. The main positives are that it is easy to interpret and implement, it can be used with unbalanced data classes and the model coefficients are easily available and can be used for understanding the importance of the variables. As main drawbacks, it prefers large datasets, it can not handle non-linearity and the model assumptions must be met for correct predictions.

3.1.2 Super Vector Machine

The SVM algorithms are very popular and often used methods for various different problems. They can both work with linear and non-linear problems.

SVMs aim to identify an n dimensional hyperplane that best divide the dataset into two classes (where n is the number of variables). Super vectors, the few data points that are considered as margins for the hyperplane are identified and then tweak the exact position and orientation.

SVMs maximize the distance between the hyperplane and the nearest data point from either class, called margin. It is a clever approach on leaving the most margin for 'errors' in the classifications.

SVMs can be applied on linearly separable data very easily. For non linear data, the kernel trick is an additional necessary step. Essentially, with the kernel trick the data gets to be transformed into a higher dimensional space where it can be separated with a hyperplane. There are various kernel types that identify and change the type of this transformation.

These algorithms are very effective on high dimensional spaces, even when the number of dimensions is greater then the number of samples. It is also a fast and memory efficient approach. Although SVM is not the best algorithm for huge datasets, it also performs poorly on noisy data and it is generally considered a black box, there is no probabilistic explanation of the produced classification.

3.1.3 Decision Tree Learning

Decision Tree learning algorithms (or DT for short) build decision trees based on an intuitive sequential decision process.

It's basic method is based on dividing nodes into at least two new nodes. This splitting is aiming to create purer sub-nodes than the parent node. The optimal split is calculated based on the criteria the method uses, that is a hyperparameter that can change the results.

Pruning is also implemented with DTs. The leaves and branches without additional explanatory power get removed from the tree. This helps reduce the model's complexity and prevents from overiftting.

DT is also a popular method, given it learns an easily understandable structure (a

decision tree). It can handle non standardized numerical data, it requires little data pre-processing. Although, DT methods are really prone to overfit, therefore small changes in the training data can influence the outcomes greatly. Using DT methods methods is often associated with long run times and can be relatively expensive. As the complexity of the problem and the data increases, the calculation times tend to become longer.

3.1.4 Random Forest

RF is an ensemble model method based on the combination of multiple different DT.

DTs are built on random subsamples of the data with different cutoffs and policies for the divisions. The combination of these smaller weaker performing models produce different, more reliable results.

The most common approach for finding the best division threshold is based on using the majority vote of the submodels, however many additional aggregations can be also implemented.

The Random Forest approach usually results in high accuracy results while being really robust to noise. It performs well on unbalanced classes and outliers do not influence the predictions greatly. Similarly to the DT methods, it can be computationally complex indicating higher training times and memory usage. Also, the interpretability provided by the DT is often lost given the more complex nature of the prediction.

3.1.5 Extreme Gradient Boosting

Similarly to RF, XGBoost is an ensemble method, that combines multiple weak performing models and outputs a more stable, better generalizing model, called strong learner.

The sequentially built submodels get weighted by giving more weight to instances where the previous models misclassify the labels. This optimization is based on the minimization of the loss function using gradient descent approach.

XGBoost shows similar advantages and disadvantages than the RF algorithm. Although, compared to RF, it is considered as a faster method, and overfitting can be avoided easier with regularization techniques.

3.1.6 Grid Search

Grid Search is a hyperparameter optimization method. It systematically iterates through a defined parameter grid, combining all combinations of hyperparameter

values. It cross-validates the training set to obtain the best hyperparameter combinations. With this approach it returns the optimal hyperparameter combination for the given model.

3.2 Sequential Pattern Mining

SPM is a data mining technique used for identifying patterns in ordered data sequences. SPM consists of discovering interesting and relevant subsequences of a set of ordered sequences. This relevancy is mostly based on the frequency or length. The sequences are ordered in a longitudinal manner, such in this work [26].

To introduce sequential pattern mining there are few key concepts that need definition. These are the following:

- **Itemset:** Collection of items.
- **Sequence:** Ordered list of itemsets.
- **Sequence database:** List of sequences, with sequence identifiers.
- **Support:** Measure of frequency of a given sequence in the sequence database.
- **Frequent Sequential Patterns:** The patterns having higher support then the minimal support that the user inputs as a cutoff.

SPM algorithms have to find all frequent patterns from the available sequence database [26]. Most of the more optimal algorithms expect a minimum support threshold and the database. Based on these two inputs they calculate the output frequent patterns.

There are numerous different approaches on how to tackle sequential pattern mining problems. The easiest and most intuitive is the naive approach where all possible sequence combinations are created, then the corresponding supports are calculated. It is obviously a very computationally heavy and inefficient approach, especially with huge sequential databases.

Multiple different approaches are available, based on different underlying algorithms. To the respect of this work VEPRECO [27], which is described in a more detailed manner given that this algorithm is used during the search for the patterns.

3.2.1 VEPRECO

The VEPRECO algorithm was proposed by Mordvanyuk, Bifet, and López [27] in 2022. The aim of the algorithm was to fasten SPM with vertical databases using pre-pruning strategies and common candidate selection.

In the works of [28] [29] [30] [31] the current most efficient sequential pattern mining algorithms CM-SPAM, SPAM, SPADE and FreeSpan are discussed. These algorithms are considered as baseline for the VEPRECO.

The VEPRECO algorithm proposed three novelties, that distinguishes it from previous solutions.

First, a new data structure is proposed based on key-value pairs instead of the previous memory-consuming data structures (vertical database).

Secondly, the pattern candidate generation in VEPRECO combines two operations (s-extensions, i-extension) into one (c-extension).

Finallly, two new pre-pruning strategies were also included for pruning the search space of the patterns. They are based on pruning the last or the last two items of the patterns.

Experiments with the baseline algorithms showed that VEPRECO is more time and memory efficient then the currently available CM-SPAM algorithm.

4 Methodology

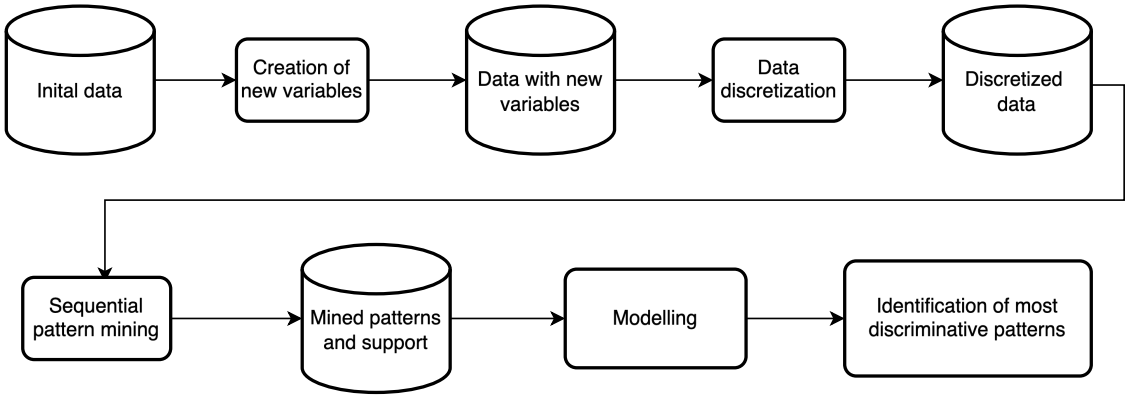

Figure 1: Methodology overview

In this chapter, the methodology to conduct the data analysis on the data, with the aim to prove the hypothesis, is described. The followed methodology is shown at Figure 1.

4.1 Data

The utilized data consisted of two cohorts of children, each collected at hospitals of Girona and Figueres (Spain). The data collection period spanned from 2008 to 2014.

The data was collected longitudinally at each doctor's visits, beginning before the birth of the observed children. Initial data pertained to the mothers and fathers during the pre-pregnancy period. Additional observations concerning the mother were conducted during the second and third trimesters. Data was also gathered at the time of birth and included the babies' measurements immediately after birth. Subsequently, data collection occurred at four intervals: 2 months, 4 months, 6 months, and 1 year after birth. The data also held the dependent variable OBESITY_B which holds information on the obesity status of the child at 5 years.

The dataset consisted of 35 variables and 386 observations. Data is imbalanced based on the dependent variable, it contains 54 Obese and 332 Not obese kids. Table 1 contains the initially collected variables with measurements and the decoding for boolean variables.

The data provided has no missing values, given that data imputation was performed in a previous work [23]. The initial dataset contained 1.304 missing values which was 9.6% of the whole dataset. Out of these, 947 missing values were numeric and 357 missing variable are categorical. In previous works data imputation was already handled, there was no need of additional imputation. The final clean version of the data consisted of 35 variables and 386 observations with imputed variables.

Variable name	Description	Units/values	Visit
AGE_M	Age of mother	years	0
HEIGHT_M	Height of mother	cm	0
SMOKING_M	Smoking status of mother (pre-preg.)	yes(1)/no(0)	0
DRINKING_M	Drinking status of mother (pre-preg.)	yes(1)/no(0)	0
BIRTH_WEIGHT_M	Birth weight of mother	kg	0
IS_FIRST_CHILD	Parity	primip.(1)/multip.(2)	0
BMI_M	BMI of mother (pre-preg.)	kg/m^2	0
AGE_F	Age of father	years	0
HEIGHT_F	Height of father	cm	0
BMI_F	BMI of father	kg/m^2	0
SMOKING_F	Smoking status of father	yes(1)/no(0)	0
DRINKING_F	Drinking status of father	yes(1)/no(0)	0
BIRTH_WEIGHT_F	Birth weight of father	kg	0
BMI_2nd_M	BMI of mother (2nd trim.)	kg/m^2	1
SYS_2nd_M	Mother's systolic BP (2nd trim.)	mmHg	1
DIA_2nd_M	Mother's diastolic BP (2nd trim.)	mmHg	1
BMI_3rd_M	BMI of mother (3rd trim.)	kg/m^2	2
SYS_3nd_M	Mother's systolic BP (3rd trim.)	mmHg	2
DIA_3nd_M	Mother's diastolic BP (3rd trim.)	mmHg	2
GEST_OBESITY_M	Gestational obesity of mother	yes(1)/no(0)	2
SMOKING_PREG_M	Smoking status of mother (preg.)	yes(1)/no(0)	2
GENDER_B	Sex of baby	girl(1)/boy(2)	3
GEST_AGE	Gestational age	weeks	3
WEIGHT_BIRTH_B	Birth weight of baby	grams	3
LENGTH_BIRTH_B	Birth length of baby	cm	3
WEIGHT_PLACENTA	Weight of placenta	grams	3
BREASTFEED	Months of breastfeeding	months	3
WEIGHT_2M_B	Weight at 2 months	grams	4
LENGTH_2M_B	Length at 2 months	cm	4
WEIGHT_4M_B	Weight at 4 months	grams	5
LENGTH_4M_B	Length at 4 months	cm	5
WEIGHT_6M_B	Weight at 6 months	grams	6
LENGTH_6M_B	Length at 6 months	cm	6
WEIGHT_12M_B	Weight at 12 months	grams	7
LENGTH_12M_B	Length at 12 months	cm	7
OBESITY_B	Obesity status at 5 years	Not obese(1)/Obese(2)	8

Table 1: Description of the collected data. The data hold information about the mother, the father and the baby at multiple time points. The descriptions of the variables are included with the corresponding units (for numerical data) or codification for boolean values. The visit column distinguishes the variables based on the date of the data collection.

4.2 Creation of New Variables

A critical point of this given project involved generating new variables from the original dataset. Previous works retained all values in their absolute forms without recording relative changes.

In this research, investigation whether using relative changes could show improved results was one of the main points. This approach with this data is generally expected to perform better, as many potential key variables are normalized. As a result, 12 new variables were computed that indicate changes in percentage (%). These values were calculated using the following formulas:

$$\text{Weight increments} = \frac{\text{Baby weight values at } t_{i+1}}{\text{Baby weight values at } t_i} - 1$$

$$\text{Length increments} = \frac{\text{Baby length values at } t_{i+1}}{\text{Baby length values at } t_i} - 1$$

$$\text{BMI mother increase} = \frac{\text{BMI of Mother at } t_{i+1}}{\text{BMI of Mother at } t_i} - 1$$

$$\text{Blood Pressure Increase (SYS and DIA)} = \frac{\text{Blood Pressure of Mother at } t_{i+1}}{\text{Blood Pressure of Mother at } t_i} - 1$$

where:

- t_i is the value of the variable at visit i.

These fractions represent the changes between two different consequent time points for each variable, indicating both the direction and magnitude of the changes in percentage points.

The newly created values are presented in Table 2. To avoid high correlations between variables, the original variables used in the calculations have been removed from the dataset.

This new updated database with increment data was used for most further steps. Due to their importance, we perform an exploratory data analysis (EDA) to understand the underlying distributions and correlations of the variables.

4.2.1 EDA of New Variables

Data distributions had been examined before discretization in relation to the prevalence of obesity in children. Plots were created for variables associated with the mother's pre-pregnancy period, the mother's pregnancy period, the father, and various variables related to the baby. The corresponding figures can be found in the appendix A.

Variable name	Description	Measures
BMI_increase_mother_2nd	BMI increase of mother (2nd trimester)	%
BMI_increase_mother_3rd	BMI increase of mother (3rd trimester)	%
SYS_BP_increase_mother_3rd	Systolic blood pressure increase of mother (3rd trimester)	%
DIA_BP_increase_mother_3rd	Diastolic blood pressure increase of mother (3rd trimester)	%
gain_length_to_2M_increment	Length gain of baby from birth to 2 months	%
gain_weight_to_2M_increment	Weight gain of baby from birth to 2 months	%
gain_length_to_4M_increment	Length gain of baby from 2 to 4 months	%
gain_weight_to_4M_increment	Weight gain of baby from 2 to 4 months	%
gain_length_to_6M_increment	Length gain of baby from 4 to 6 months	%
gain_weight_to_6M_increment	Weight gain of baby from 4 to 6 months	%
gain_length_to_12M_increment	Length gain of baby from 6 to 12 months	%
gain_weight_to_12M_increment	Weight gain of baby from 6 to 12 months	%

Table 2: Description of the added new variables.

Based on these plots it is possible to observe that the distribution of binary values is quite uneven. This is mostly true in case of the variables GEST_OBESITY_M, DRINKING_M, SMOKING_M and SMOKING_PREG_M.

The correlations of the variables were calculated based on the Pearson correlation method, that quantifies the linear relationship within variables. A correlation plot of the calculated correlation of the variables is shown in Figure 16.

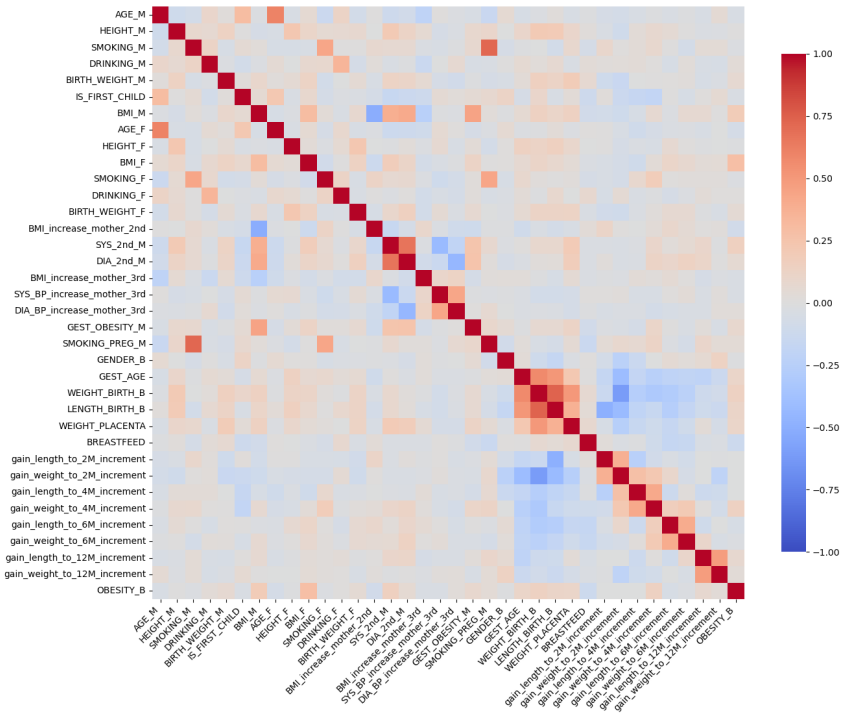

Figure 2: Correlation of the variables

The plot reveals that, overall, there is no strong correlation between the variables. The highest positive correlations are observed between the smoking status before and after pregnancy, the pre-pregnancy diastolic blood pressure and its change, as well as the baby's birth weight and birth length. The highest negative correlations are found

between the mother's initial BMI and the BMI increase to the second trimester, and between the baby's birth weight and the second weight gain increment.

4.3 Data Discretization

Data discretization is the process of converting continuous numerical data into discrete, categorical bins. Implementing this process was important to create suitable data for the SPM algorithm. Discretization simplifies the data by reducing the number of unique values, therefore enabling the algorithm to identify reliable patterns more effectively.

During the discretization of the data, three guidelines were followed:

- **Guideline 1:** Preserve as much of the added original information as possible.
- **Guideline 2:** Ensure each variable has the same number of categories, except the ones where the added original information makes more necessary.
- **Guideline 3:** Achieve a balanced number of observations in each category.

Following these defined guidelines, the variables were categorized into three groups. The binning of the variables were handled identically within each group. The groups were the following:

- **Group 1:** Numerical variables with added information of their values (BMI of the mother and father, gestational age, blood pressure values)
- **Group 2:** Numerical variables without added information of their values (e.g.: age of mother and father, height of mother and father, increments)
- **Group 3:** Categorical variables.

Up to 5 variables were associated with group 1. These were BMI of the mother and the father, gestational age, systolic and diastolic blood pressure. The cutoff values for the bins were based on WHO recommendations for these variables. All recommendations divided the variables into 4 distinct groups except the BMI. To keep the added meaning behind the BMI values, 5 categories were necessary (see Guideline 2.).

All other numerical variables belonged to group 2. The cutoff values for these variables were at the quartiles of the observations, therefore all of these numerical values were categorized into 4 groups.

The initially categorical values were kept were kept with the same number of values.

After discretizing the variables, unique codes were assigned to each category within each variable.

For the first variable (AGE_M), the categories were coded as 1, 2, 3, and 4, each representing the corresponding quartiles of the mother's age. For the second variable (HEIGHT_M), the categories were coded as 11, 12, 13, and 14. For the third variable (SMOKING_M), the categories were coded as 21 (non-smoking) and 22 (smoking), and this pattern continued up to the last variable, which had categories coded as 341, 342, 343, and 344.

The structure of the codes was such that the first two digits represented the variable, while the second digit represented the category within that variable. The codes incremented by 10 for each variable, ensuring that each variable had a distinct range of codes.

This coding scheme provides clarity by clearly differentiating between variables and categories; for instance, code 12 represents the second category of the first variable, and code 32 represents the second category of the third variable. The systematic approach makes it easier to identify which variable and category a specific code belongs to.

The created categories, cutoff values and the codifications are included in the Appendix B.

4.3.1 EDA of Discetized Data

Due to the nature of the discrete data, statistic independence testing was applied to examine the prevalence of associative relationships. A Chi-squared test was performed to determine if there were significant relationships between the variables. A relationship between variables was considered present if the p-values from the test fell below 0.05. With checking all combinations a simplified heatmap was created with only the significant associations, shown at Figure 3.

The major limitation of using Chi-squared testing is that, although it can reveal the presence of a significant relationship, it does not provide information about the magnitude or direction of that relationship. Consequently, while the existing relationships are observed the extent of their influence can not be determined on the results or subsequent work.

4.4 Sequential Pattern Mining

Using discretized variables and making the necessary formatting adjustments, the pattern mining process was initiated. This pattern mining was conducted twice, once for the Obese and once for the Not obese children, to identify different factors based on the class of the dependent variable. This step was necessary to determine

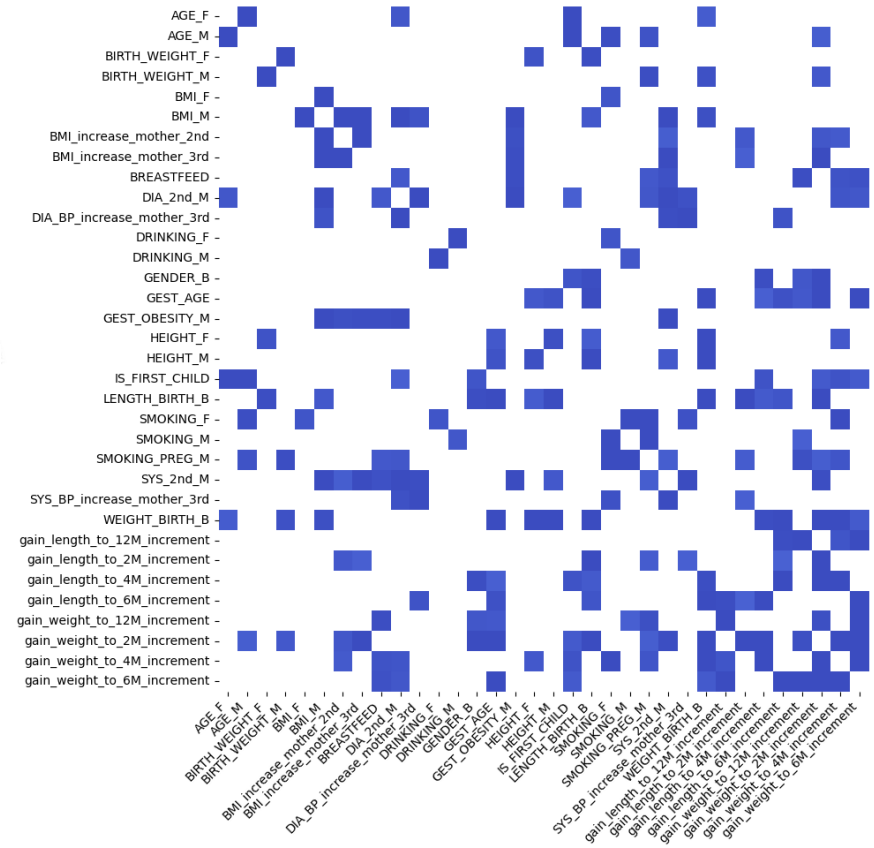

Figure 3: Significant association relationships between variables

the distinct differentiating influencing factors for the Obese and Not obese classes.

To optimize the VEPRECO algorithm, various parameter settings were experimented with. The goal was to mine sequences of any length with varying gaps between consecutive elements while obtaining a reasonable number of detected patterns.

The min_sup_rel parameter, representing the minimum support relative to the total number of sequences, was adjusted to influence the number of detected patterns. After testing different values heuristically, a value of 0.4 was set. With this min_sup_rel value, 205 patterns were mined for the Obese class and 231 for the Not obese class. After taking the symmetric difference and removing common patterns between the classes, 150 distinct patterns related to either the Obese or Non obese class were identified.

With the mined patterns, a dataset was created to represent which initial observations held the given patterns and what the dependent variables were at the observation. Using this dataframe, experiments with modeling were initiated. An example representation of the final database can be seen in Figure 4.

Observation	Pattern 0	Pattern 1	Pattern 2	Pattern 3	...	Pattern 149	OBESITY_B
0	0	0	1	1	...	1	1
1	0	0	0	0	...	0	0
2	1	0	1	1	...	0	0
...	...	...	...	...	...	...	...
385	0	0	0	0	...	0	1

Figure 4: Example of the created dataframe. The data holds information about the 150 patterns and the dependent variable, obesity status.

4.5 Modelling

A predictive model has been built with the new data using up to five machine learning methods, in a two sequential steps.

In the first step all five methods have been used without finding and employing the best possible hyperparameteres, with the aim of selecting the three best suiting algortihms for the data.

In the second step, grid search has been conducted with previously chosen three methods, in order to learn the best model. Using the extracted best hyprparameters, new models were trained. Out of these optimized versions of the models the most relevant patterns were extracted and analyzed.

4.6 Identifying the Factors of Obesity

Identifying the obesity factors was the final step.This involved determining the significant variables and model coefficients, along with their corresponding p-values, using LR's individual Wald test. For DT and RF feature importances were evaluated.

Additionally, Shapley Additive Explanation values (SHAP) were calculated and visualized to provide further insights. The obtained patterns were examined and contrasted against clinically significant factors.

4.7 Experimental Setup

To test the used methodology, three different experiments have been defined. These are the following:

- **Experiment 1. - Main experiment:** Finding the optimal obesity detection algorithm and evaluate the obtained patterns.
- **Experiment 2.:** Comparison of results obtained in sequential pattern mining and itemset mining.
- **Experiment 3.:** Comparison of results obtained with or without increments.

4.7.1 Experiment 1. Model Selection and Pattern Identification

In the first modeling experiment focused on finding the most optimal predictive model for the classification task. During the experiment five different models were evaluated: Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and XGBClassifier. To determine the best-performing model, stratified K-Fold cross-validation with 5 folds was conducted, testing each model on the entire dataset.

The evaluation criteria included accuracy, F1 score, precision, and recall. Given the goal of shedding light on factors contributing to obesity, optimizing recall was prioritized. As recall measures the ability to correctly identify Obese individuals among all true Obese cases, recall was crucial for identifying factors contributing to obesity. Additionally, accuracy was considered to measure overall model performance.

Based on the cross-validation results, the three best-performing models were selected: Random Forest, Decision Tree, and Logistic Regression. The models were fine-tuned to optimize their classification results. This involved performing grid searches, with a focus on optimizing recall, to find the best combination of hyperparameters for each model.

4.7.2 Experiment 2. Itemset versus SPM

In the second experiment, the obtained results from itemset mining [23] were investigated and contrasted against the results found with SPM in this work.

Experiment 2 comprised two sub-experiments. In the first sub-experiment, the top 200 itemset patterns (based on relevance) were compared against all patterns discovered through sequential pattern mining. In the second sub-experiment, the top 20 most relevant patterns from both methods, specifically those extracted using a DT algorithm, were identified and compared with each other.

The two methods were compared with cross validation, on a DT classifier. They were contrasted based on the recall and accuracy scores.

4.7.3 Experiment 3. Value Added by Incremental Data

As the inclusion of increments was a key novelty in this work compared to previous works, experiment 3. was conducted to assess the added value of including increment data. In a [23], DT classification was performed using the original, non-discretized data. To evaluate the impact of including increment data, the same modeling experiment was replicated, this time incorporating the increment data, to determine its effect on model performance.

The testing consisted if 5 fold cross validation. The corresponding accuracy and recall scores were extracted to calculate and plot the scores.

At experiment 3, the DT was employed, once with the default hyperparameters, which were used during previous works, and once with the ones this study found most optimal, described at the section 5.3.1.

5 Results

In this chapter the obtained results are described. First sections detail the results of the process for finding the best predictive models, and next, detailed results of the hyperparameter optimization and patterns found by the best models.

Section 5.5 describes the results of the use of itemset mining versus SPM, while section 5.6 describes the results obtained using the row data or with increments.

The discussion of the extracted results are described in more detail are provided in the next chapter.

5.1 Model Selection

Initial experimentation with using K-Fold Cross validation was concluded in order to find the best approaches for further modelling. All the available dataset that held the pattern information was used. 5 possible models were tried, Logistic Regression, Decision Tree, Random Forest, XGBoost and Super Vector Machine classifier.

Visualizations were created based on the used metrics during the evaluation of the models. Figure 5 shows the plots of how the given metric changed after adding new patterns iteratively.

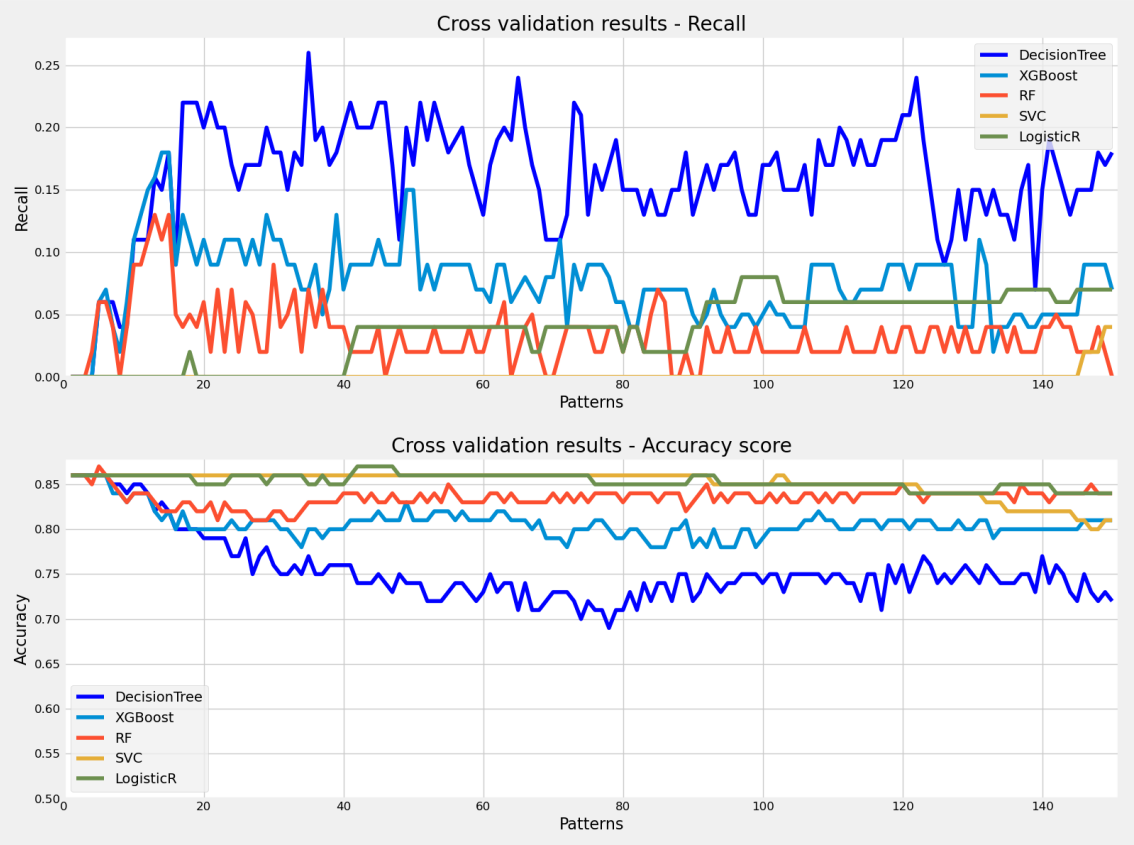

Figure 5: Recall and accuracy scores on CV

Based on Figure 5. RF and LR achieved the highest accuracy with 150 patterns, while DT and LR performed best in terms of recall. Based on these results, further experiments were conducted using these three models: RF, DT, and LR.

5.2 Logistic Regression

As LR is one of the methods that perform the best, with the default hyperparameters, it has been selected for additional work, to find which factors identify as key for child obesity.

5.2.1 Grid Search on Hyperparameters

The preformed grid search was optimized on the recall score. The optimal hyperparameters were the following seen below, while the used random state was 42:

- **C:** 100
- **Penalty:** l1
- **Solver:** liblinear

5.2.2 Results

The training achieved an accuracy of 71%. For the Obese group, the recall was 0.12, while for the Not obese group, it was 0.81. The precision scores were also similar to the recall results, with the Not obese class having a precision of 0.84 and the Obese class having a precision of 0.1.

5.2.3 Obtaining the Most Important and Significant Patterns

Once the logistic model was trained, the most important patterns used for the predictions could be extracted. To get these values, the absolute coefficients of the model were calculated. The patterns associated with these values had the highest importance on the model.

Once these values were available, the p-values for the individual Wald tests of the variables (patterns) were calculated. The obtained relatively low p values indicated that the model had many relevant variables. In order to extract the ones with the highest importance the top 20 features were selected based on absolute highest coefficients with p values below 0.05.

The extracted patterns were decoded in order to make interpretation easier. The list of extracted patterns can be seen at Table 3. The table shows the most important patterns in descending order by the absolute value of the coefficients. It also shows

the decoded meaning behind the important patterns and the classes that the given pattern is associated with.

N. Pattern	Pattern	Class
12	BIRTH_WEIGHT_F=3, SYS_2nd_M<115, GEST_OBESITY_M=0	Not obese
106	25<=BMI_F<30, DRINKING_M=0	Obese
32	SYS_2nd_M<115, 18.5<=BMI_M<25	Not obese
100	SMOKING_PREG_M=0, SYS_2nd_M<115, 18.5<=BMI_M<25	Not obese
19	38<=GEST_AGE, BIRTH_WEIGHT_M=3, SMOKING_F=0	Obese
88	38<=GEST_AGE, BIRTH_WEIGHT_F=3, 18.5<=BMI_M<25	Not obese
133	DRINKING_F=1, BIRTH_WEIGHT_F=3, BIRTH_WEIGHT_M=3	Obese
22	IS_FIRST_CHILD=1, GEST_OBESITY_M=0	Not obese
57	DRINKING_F=1, SMOKING_PREG_M=0, BIRTH_WEIGHT_M=3	Obese
40	BIRTH_WEIGHT_M=3, 25<=BMI_F<30	Obese
119	SMOKING_PREG_M=0, BIRTH_WEIGHT_M=3, 18.5<=BMI_M<25	Not obese
23	DRINKING_F=1, 38<=GEST_AGE, BIRTH_WEIGHT_F=3	Obese
108	IS_FIRST_CHILD=2, GEST_OBESITY_M=0	Not obese
140	38<=GEST_AGE, GEST_OBESITY_M=0, SMOKING_PREG_M=0, DRINKING_M=0	Not obese
30	38<=GEST_AGE, SMOKING_PREG_M=0, BIRTH_WEIGHT_M=3, SMOKING_F=0	Obese
83	38<=GEST_AGE, SYS_2nd_M<115	Not obese
1	BIRTH_WEIGHT_M=3, 18.5<=BMI_M<25, GEST_OBESITY_M=0	Not obese
67	SMOKING_PREG_M=0, BIRTH_WEIGHT_M=3, 25<=BMI_F<30	Obese
116	38<=GEST_AGE, SMOKING_F=0, SMOKING_M=0, DRINKING_M=0	Obese
89	SMOKING_PREG_M=0, IS_FIRST_CHILD=2	Not obese

Table 3: Results table for Logistic Regression.

5.2.4 SHAP Values Results

The SHAP values were calculated based using the SHAP library [32]. The obtained scores on the LR could be found in Figure 6.

Throughout the work, feature importance scores were primarily used to evaluate results. However, SHAP values offer deeper insights into the underlying significance of these scores. It's important to note that 18 patterns were identified with both methods. This suggests that the top 20 extracted patterns are robust and consistent, as they are corroborated by two distinct evaluation techniques.

5.3 Decision Trees

DT is another of the methods that performs the best with the default hyperparameteres. Therefore, additional work has been conducted in order to find the relevant factors.

5.3.1 Grid Search on Hyperparameters

Once again, during the optimization with grid searching, the aim was to obtain the highest possible recall score. The obtained hyperparameters for the DT algorithm were the following:

- **Criterion:** Gini

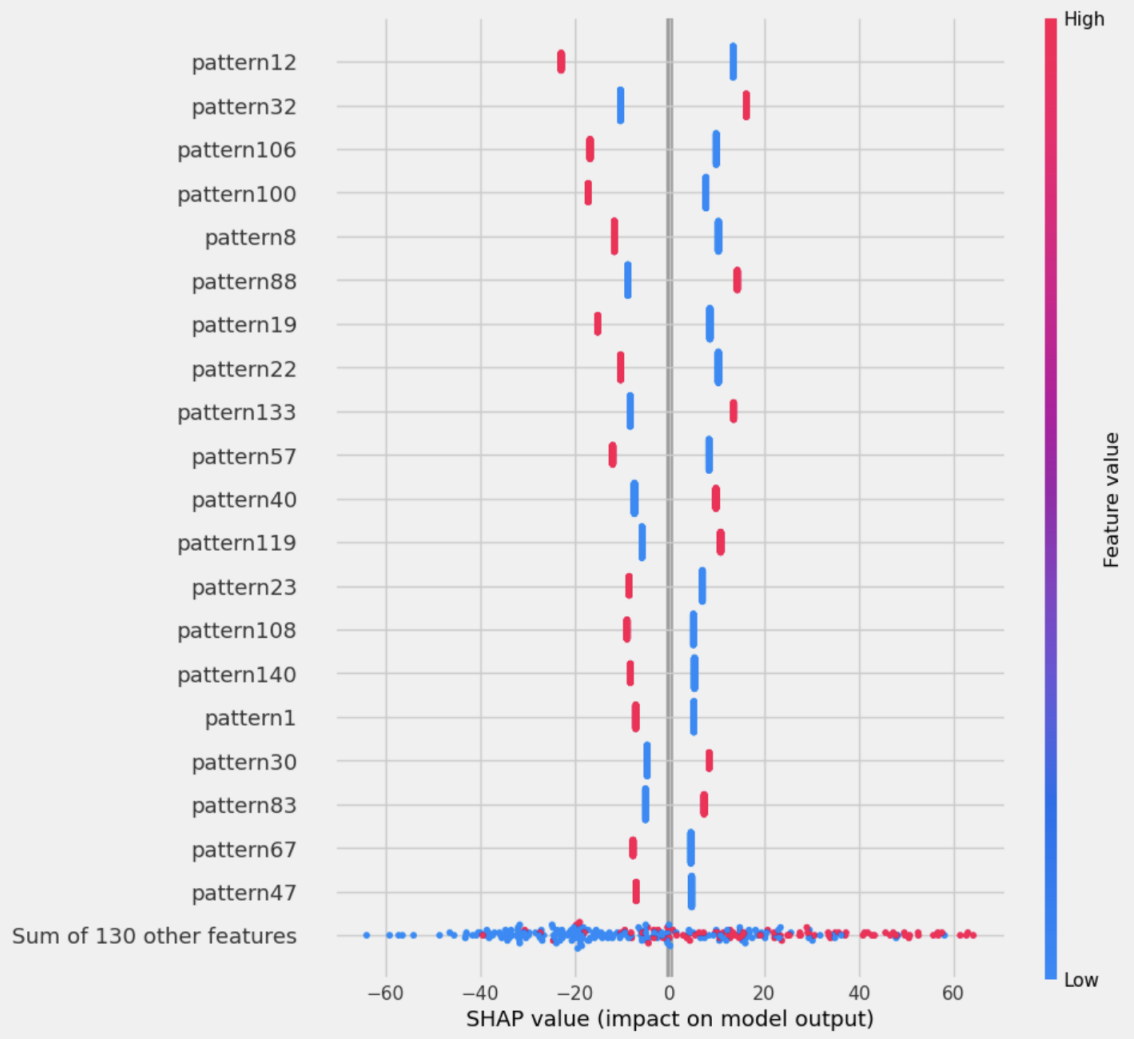

Figure 6: SHAP values of Logistic Regression classifier

- **Maximum depth:** None
- **Minimum samples per leaf:** 1
- **Minimum samples per split:** 5

5.3.2 Results

The previosuly described hyperparameters were utilized for the DT classifier. The model produced results that closely matched those of the LR mentioned previously, yielding a slightly improved accuracy of 72%. Specifically, the recall for the Not obese class was 0.82, while for the Obese class, it was 0.12. Similarly, the precision scores were comparable, with values of 0.84 and 0.1, respectively.

5.3.3 Feature Importance Results

To obtain the most significant patterns feature importance method has been applied. Once the results were obtained the patterns having 0 importance were removed. Out of the remaining 32 features were chosen the top 20 features that had the highest importance can be seen in Table 4.

N. Pattern	Description	Class
34	38<=GEST_AGE, 3000<WEIGHT_BIRTH_B<=3500	Not obese
149	0.06<=gain_length_to_6M_increment<0.08	Not obese
123	0.1<=gain_length_to_12M_increment<0.125, GEST_OBESITY_M==0	Not obese
107	38<=GEST_AGE, SMOKING_PREG_M==0, 25<=BMI_F<30	Obese
39	0.1<=gain_length_to_4M_increment<0.125	Obese
35	SMOKING_PREG_M==0, BIRTH_WEIGHT_F==3, 18.5<=BMI_M<25, GEST_OBESITY_M==0	Not obese
27	SYS_2nd_M<115, 18.5<=BMI_M<25, GEST_OBESITY_M==0	Not obese
17	38<=GEST_AGE, SYS_2nd_M<115, GEST_OBESITY_M==0	Not obese
114	SMOKING_PREG_M==0, 25<=BMI_F<30, SMOKING_M==0	Obese
148	38<=GEST_AGE, BIRTH_WEIGHT_M==3, SMOKING_F==0, GEST_OBESITY_M==0	Obese
73	3000<WEIGHT_BIRTH_B<=3500, BIRTH_WEIGHT_F==3	Not obese
83	38<=GEST_AGE, SYS_2nd_M<115	Not obese
144	0.15<=gain_weight_to_6M_increment<0.2	Obese
26	0.125<=gain_length_to_2M_increment<0.175, SMOKING_M==0	Not obese
49	GEST_OBESITY_M==0, BIRTH_WEIGHT_F==3, SMOKING_M==0, DRINKING_M==0	Not obese
76	IS_FIRST_CHILD==1, SMOKING_M==0	Obese
64	0.125<=gain_length_to_2M_increment<0.175, SMOKING_PREG_M==0, SMOKING_M==0	Not obese
36	SMOKING_PREG_M==0, BIRTH_WEIGHT_M==3, SMOKING_F==0, DRINKING_M==0	Obese
111	DRINKING_F==1, 18.5<=BMI_M<25	Not obese
15	BIRTH_WEIGHT_M==3, GENDER_B==1	Obese

Table 4: Results table for Decision Tree classifier.

5.3.4 SHAP Values Results

The extracted results could be seen in Figure 7. The image shows the top 20 patterns based on the SHAP values.

Both methods of pattern extraction identified 15 overlapping features in their top 20 lists, based on feature importance and SHAP values. These patterns represent features that are consistently highlighted as influential by both metrics.

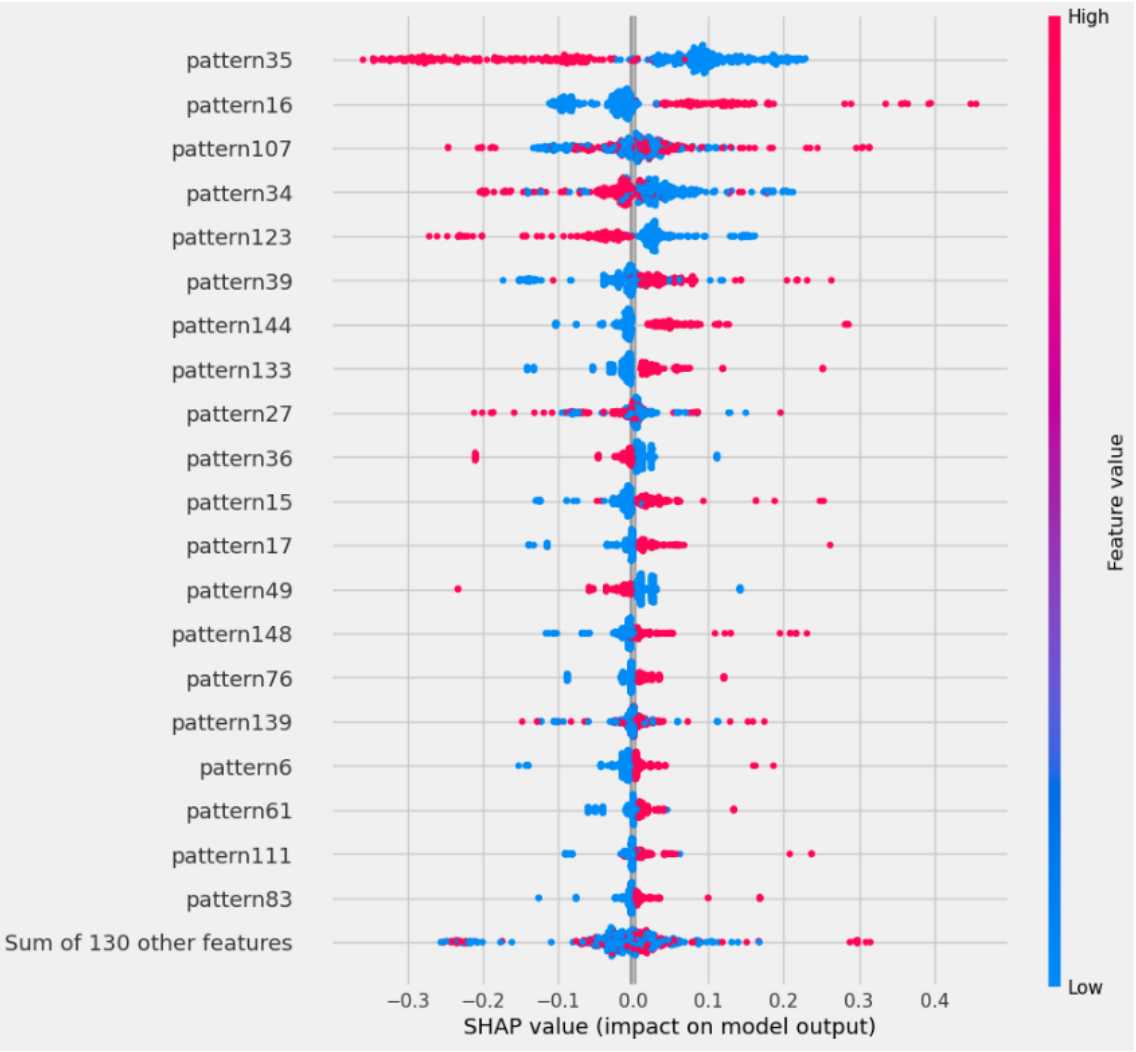

Figure 7: SHAP values of Decision Tree classifier

5.4 Random Forest

RF is the third method selected in the first experiment. In this section, the results on a deeper work on this method are provided.

5.4.1 Grid Search on Hyperparameters

Just as previously after optimizing my grid search on recall, the following hyperparameters have been found:

- **Maximum depth:** None
- **Maximum features:** Auto
- **Minimum samples per leaf:** 2
- **Minimum samples per split:** 2
- **Number of estimators:** 100

5.4.2 Results

While achieving the highest accuracy of 83% with both models using RF classifier, the results for recall and precision were less satisfactory. Specifically, the recall for the Not obese class was 0.97, indicating strong performance in identifying Not obese individuals. However, the recall for the Obese class was 0, suggesting that the model fails to identify any instances of obesity. Similarly, precision scores were 0.85 for the Not obese class and 0 for the Obese class, indicating that while the model identifies Not obese individuals fairly well, it struggles to correctly classify Obese individuals, often predicting them as Not obese.

5.4.3 Feature Importance Results

All 150 patterns contributed to obesity prediction, though often with values very close to zero, indicating that they have little to no explanatory power in the particular model.

Next, the top 20 patterns based on their importance values were selected. These variables are presented in Table 5.

5.4.4 SHAP Values Results

SHAP values were computed for the RF, as previously done. The results of these calculations are shown in Figure 8.

Upon reviewing the patterns, there were only 8 common patterns between the top 20 feature importances and SHAP values.

N. Pattern	Description	Class
72	SMOKING_PREG_M==0, 0.1<=gain_length_to_4M_increment<0.125	Obese
144	0.15<=gain_weight_to_6M_increment<0.2	Obese
133	DRINKING_F==1, BIRTH_WEIGHT_F==3, BIRTH_WEIGHT_M==3	Obese
117	0.1<=gain_length_to_12M_increment<0.125	Not obese
149	0.06<=gain_length_to_6M_increment<0.08	Not obese
39	0.1<=gain_length_to_4M_increment<0.125	Obese
71	3000<WEIGHT_BIRTH_B<=3500, GEST_OBESITY_M==0	Not obese
139	GENDER_B==0, GEST_OBESITY_M==0	Not obese
47	IS_FIRST_CHILD==1, BIRTH_WEIGHT_M==3	Obese
64	0.125<=gain_length_to_2M_increment<0.175, SMOKING_PREG_M==0, SMOKING_M==0	Not obese
68	38<=GEST_AGE, BIRTH_WEIGHT_M==3, SMOKING_M==0, DRINKING_M==0	Obese
127	DRINKING_F==1, 38<=GEST_AGE, BIRTH_WEIGHT_M==3	Obese
120	3000<WEIGHT_BIRTH_B<=3500, SMOKING_PREG_M==0	Not obese
130	SMOKING_PREG_M==0, GENDER_B==0	Not obese
96	38<=GEST_AGE, GENDER_B==1	Obese
6	SMOKING_PREG_M==0, IS_FIRST_CHILD==1, SMOKING_M==0	Obese
16	BIRTH_WEIGHT_M==3, SMOKING_M==0, GENDER_B==1	Obese
60	60<=DIA_2nd_M<70, BIRTH_WEIGHT_M==3	Obese
66	SMOKING_PREG_M==0, 0.1<=gain_length_to_12M_increment<0.125	Not obese
90	BIRTH_WEIGHT_F==3, SYS_2nd_M<115	Not obese

Table 5: Results table for Random Forest classifier.

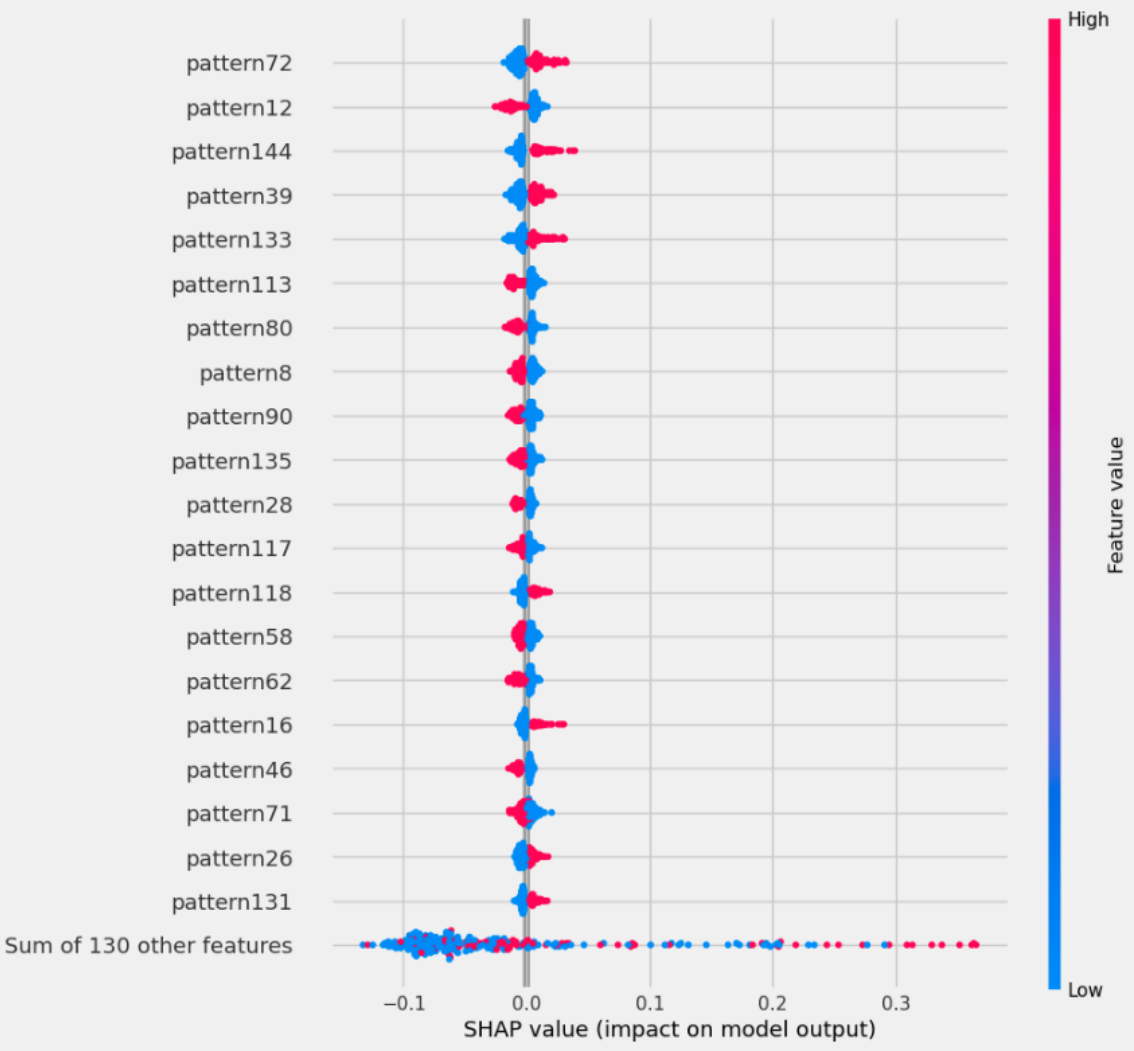

Figure 8: SHAP values of RF

5.5 Itemset Mining versus SPM

As described at section 4.6.2, two experiments were conducted in order to find the differences in the results of itemset mining and sequential pattern mining. One subexperiment was focused on how large number of extracted patterns performed on Decision Tree classifiers. All 150 mined sequential pattern mining patterns were used alongside with the 200 most relevant itemset patterns. The results of the comparison can be found at figure 9.

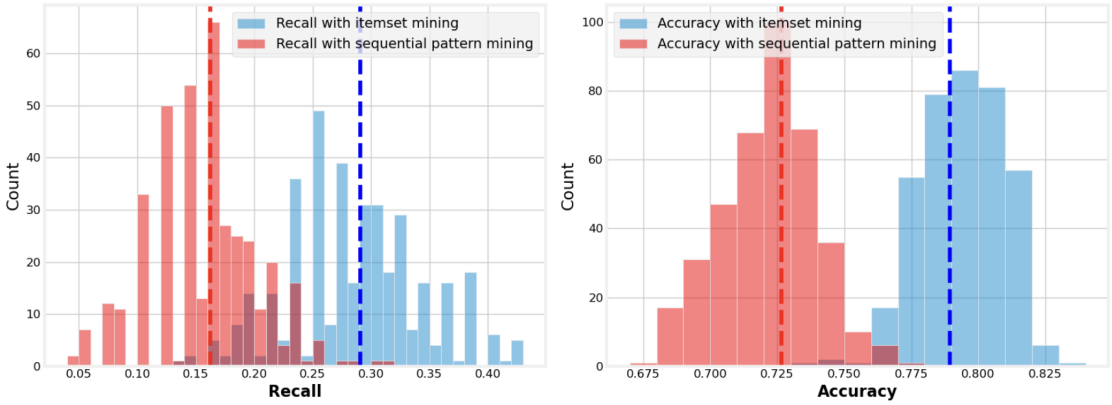

Figure 9: Recall and accuracy values extracted when using all sequential pattern mining patterns and top 200 itemset mining patterns

The second sub experiment was based on comparing the top 20 patterns extracted from each method. The obtained plot can be found in Figure 10.

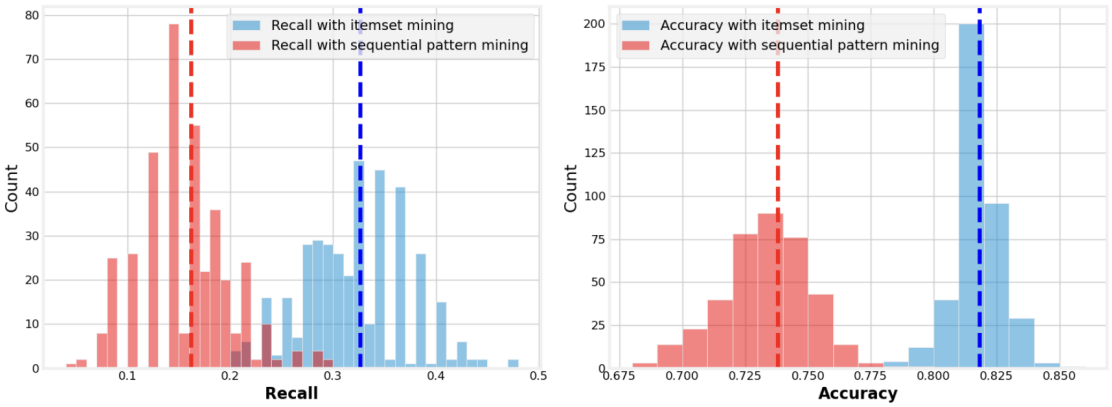

Figure 10: Recall and accuracy values extracted when using all top 20 sequential pattern mining pattern s and top 20 itemset mining patterns

5.6 Value Added of Incremental Data

Accuracy and recall scores were obtained from two experiments with Decision Tree classification on the original data and the data with the included increments. The resulting plots can be found are in Figures 11 and 12.

The plots show the distribution of the obtained scores at all cross validation instances. The dotted lines stand for the average scores obtained from the given measure.

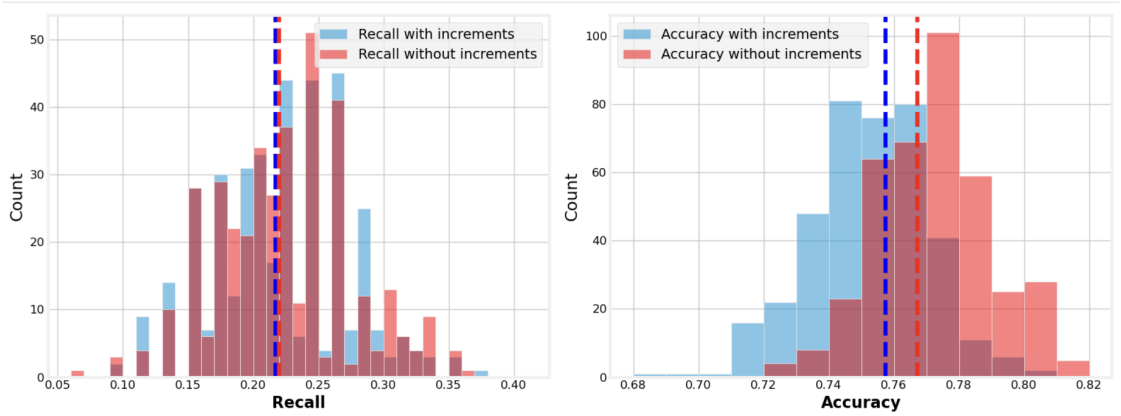

Figure 11: Accuracy and recall scores of the original data and the data with increment values. The modelling was conducted with DT with the default hyperparameters.

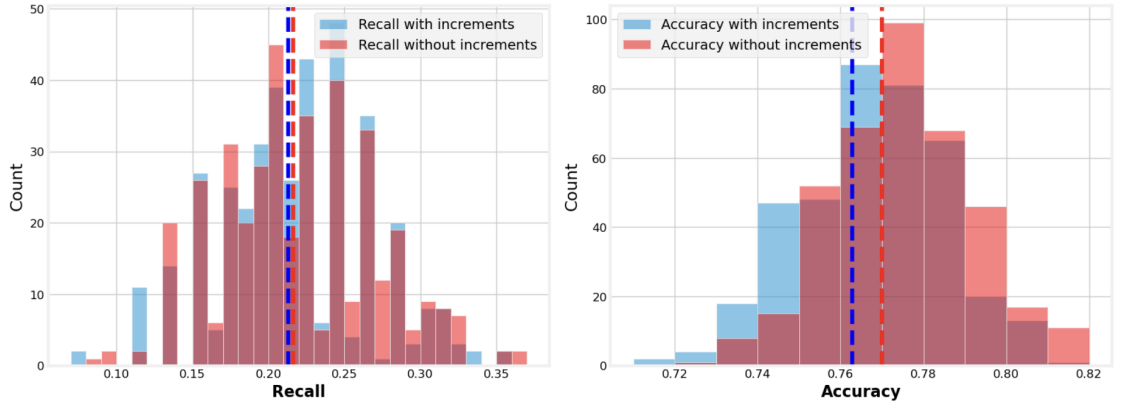

Figure 12: Accuracy and recall scores of the original data and the data with increment values. The modelling was conducted with DT with the hyperparameters obtained at grid search.

6 Discussion of the Results

This chapter first presents the results of the experiment 1, focusing on the patterns observed within the two classes and the commonalities across various methods. The findings are compared with existing scientific and clinical research on childhood obesity factors. Later it discusses the results obtained from experimental case 2 and experimental case 3.

6.1 Logistic Regression Results

The LR performed the worst accuracy score, while providing reasonable recalls and precision even for the underrepresented Obese class. This indicates that even though it might not be the strongest performing model, it is still able to detect both Obese and Not obese classes correctly in some cases.

6.1.1 Found Patterns - Obese Class

The model detected 9 Obese patterns from the top 20 variables by absolute value of coefficients.

Parental variables were present in all 9 patterns, with the most prevalent being a positive paternal drinking status, observed in 3 of the patterns. A combination of higher birth weights or BMI of the parents was also observed in these patterns.

The smoking statuses of both mothers and fathers were detected multiple times, with negative outcomes (nonsmokers) in every case. These results contradict the initial expectations.

Higher body mass indexes or birth weights were frequently observed on both the maternal and paternal sides. This aligns with the currently available information on causality between higher parental weights, birth weights and the children obesity.

Lastly, low gestational age (under 38 weeks) was detected in 3 patterns. In all instances, this was combined with either negative smoking or negative maternal drinking statuses, or with high maternal birth weight.

6.1.2 Found Patterns - Not Obese Class

Low systolic blood pressure at the second trimester (below 115) was observed multiple times in the relevant patterns. Most of the cases this was combined with normal BMI indexes of the mothers (between 18.5 and 25).

Negative gestational obesity was detected multiple times. It was combined with both primiparous and multiparous pregnancies, indicating that parity is a less important factor.

Surprisingly low gestational age was detected multiple times. In all cases low gestational age was combined with relevant features, such as no gestational obesity, negatives parental smoking status or healthy BMI scores.

6.2 Decision Tree Results

The results obtained with the DT were slightly better. A 0.72 accuracy value with higher recall and precision scores of the Obese class has been obtained. This model performed best at detecting the Obese cases.

6.2.1 Found Patterns - Obese Class

Similarly to the results found at Logistic Regression, parental BMI and parental birth weight seem to be highly prevalent in the patterns. In almost all cases these variables are combined with low gestational age and negative parental smoking.

The detected smoking status was negative in all cases. Again, this contradicts the expectations based on current research.

Variables detected with the DT are the increments of weight and length growth of the baby. The patterns indicate that higher growth values at 4 and 6 might lead to obesity.

6.2.2 Found Patterns - Not Obese Class

The weight and length increments were present at the Not obese class patterns as well. Most of these increment values were associated to the second quartiles. Indicating that non extreme growth values, that can be considered , are associated with the Not obese class. These growth values were combined with nonsmoking parents and normal BMI and birth weight values.

Low gestational age values were identified many instances which also contradict expectations. Although all of the values were combined with otherwise expected variables such as normal systolic blood pressure values and health BMI indexes.

6.3 Random Forest Results

RF resulted with the best accuracy score, by far over-performing the other two models. With that being said, recall and precision scores on the Obese class seriously undermine the results obtained by the other algorithms. Due this, the detected Obese class patterns are seriously questionable.

6.3.1 Found Patterns - Obese Class

Similarly to the DT results third quartile growth increments at 4 and 6 months were detected. These variables were combined with higher BMI and paternal birth weight or with negative smoking pregnancy.

The fathers' positive drinking status was detected at multiple patterns. This variable was combined with higher maternal birth weights at both cases, while low gestational age and high paternal birth weights were also prevalent.

The gestational age was once again represented in the patterns. It was combined the previously seen variables such birth weights and non-smoking, non-drinking statuses.

One additional new variable was present in the patterns, which was the gender of the baby. In both cases when the variable was mined, the sex of the baby was girl. This variable was combined, again with the previously seen variables, low gestational age, maternal non-smoking and high maternal birth weight.

6.3.2 Found Patterns - Not Obese Class

High number of increment data was included in the patterns. All of these were related to the length measurements. The relevant patterns were associated to the 2nd, 6th and 12th months. All the mined variables were detected at the second quartile, the lower-normal end of the length growth.

In contrast to the Obese class, the gender of the baby detected in the Not obese patterns was boys. It was combined with non smoking of the mother.

The birth weight of the observed babies detected weights between 3-3.5kg at 2 instances. These were combined with nonsmoking and gestational obesity.

Negative smoking statuses were detected again, at multiple patterns.

6.4 Overlapping Patterns

There were 6 patterns that appeared across multiple model approaches. The highest overlap occurred between the DT and RF classifiers, which is expected considering their similar underlying logic and methodology. Specifically, patterns 39, 64, 144, and 149 were common between these two models.

When combining the results of LR with other models, there were fewer overlaps. Each model combination had only one overlap: pattern 83 overlapped with DT, and pattern 133 overlapped with RF.

There was no pattern that was prevalent in both three model approaches.

Among the 6 overlapping patterns, three were associated to the Not obese class and three were associated to the Obese class.

The Obese class patterns were prevalent at the combinations of DT - RF and the RF - LR models. These Obese patterns were associated with length and weight increments, both related to 3rd quartile values. One, clinically suitable pattern, positive drinking status of the father, higher birth weight of the father and the mother was also detected. All of these overlapping patterns align with the expectations.

The Not obese class patterns are found at the DT - LR and the DT - RF model combinations. Here the detected patterns were also associated with length gain values, one at the second quartile (at 2 months and one at the third quartile (6 months). The 2 month length development variable was also combined with negative maternal smoking statuses (pre-pregnancy, during pregnancy). Low systolic blood pressure was detected with low gestational age, which contradicts the expectations.

6.5 Summary of the Identified Patterns

Most models found similar deterministic variables and similar sequential patterns that were used during the classification. The length and weight increment variables seem to divide the classes fairly well. Most cases third quartile (higher growth) increments were related to obesity while second quartile was associated with Not obese class.

The parental BMI, parental birth weight, gestational obesity variabes align with the expectations, and were detected multiple times in the patterns detected by all models. This might indicate the importance of these factors.

Surprisingly, varibles parental smoking and low gestational age do not align with our expectations. These variables were detected in multiple patterns at both classes. The reason of this frequent detection might be lying in the data's structure. It seems that both these variables were very unbalanced and that is why the pattern mining algorithm detected them so often.

There were also multiple variables that were only present at one of the models. These included, sex of the baby, parity and diastolic blood pressure.

The longitudinal aspect of the data did not became as prominent as expected. The obtained patterns did not show strictly related patterns, such as weight or length changes or BMI changes. from visit to visits. In most cases the variables in the found patterns are somewhat unrelated to the timely manner.

6.6 Itemsets versus Sequence Mining

The conducted experiments showed that in both cases the results of itemset mining outperformed the results of SPM. Both in terms of accuracy and recall, the patterns extracted with SPM were seriously under-performing.

The dotted lines in Figures of Chapter 5, are representing the average of the obtained scores. The lines show significant differences alongside with the distributions of the obtained measures. Based on these results it is apparent that itemset mining is the more optimal method for obtaining the patterns from the utilized data.

6.7 Value Added from Incremental Data

The two experiments with the original and the increments data contradict the initial expectations. It is apparent that using the increments instead of the absolute results impacts negatively on the model performance.

Both tweaks of the algorithm (default and non-default hyperparameters) showed that the models with absolute data performed better in terms of accuracy and recall. Both cases the mean of the metrics were higher with the original data.

Based on these results, we can state that incorporating the increments did not lead to better results.

7 Limitations and Future Work

This chapter describes the limitations faced during the work alongside the possible future work related to this study.

7.1 Limitations

The limitations of the presented work are notable in several respects. Firstly, the major constraint relates to the dataset itself, consisting of only 386 observations. This sample size is relatively small for achieving optimal results in machine learning applications.

Additionally, the initial dataset shows imbalance, affecting both the explanatory and dependent variables. This imbalance significantly influences the outcomes. For example, disparities in smoking statuses are prevalent across many patterns irrelevant to the obesity status. Based on this it is deductible that they contribute little to no additional explanatory power.

Moreover, the imbalance in the dependent variable, the obesity status undermines model efficiency, as it reduces the amount of data from which models can learn. Consequently, all models perform poorly in accurately predicting the Obese class.

The study employed a limited range of methods. While employing more advanced AI modeling techniques such as neural networks could have been beneficial, the small number of observations is sub-optimal for such experiments.

7.2 Future Work

The previously described limitations highlight the challenges and potentials for further exploration in future research. To tackle the main challenge of having this low number of observations it is highly recommended to collect more observations or create synthetic data based on the currently available dataset.

Exclusion of highly unbalanced variables is also recommended. This includes the smoking and drinking statuses and the gestational age variables.

Given the negative results obtained with the employment of increment data, it might be advised to test the methods with the original absolute data.

There are additional possible relevant variables to be included in further research. These could include the babies status, whether they are too small or too large for gestational age. Both are considered as risk factors for later life obesity and could be calculated from the currently available data [33] [34].

8 Conclusion

This presented study focused on extracting the most important factors leading to childhood obesity based on Catalan hospitals longitudinal data. The key novelty presented in this work alongside the methods utilized were the initial data preprocessing step of creating increment data instead of using absolute values and a representation of instances based on sequence patterns.

A sequential pattern mining algorithm was utilized to extract the most prevalent patterns. The algorithm extracted 150 distinct patterns associated with either Obese or Not obese children at five years old. These extracted patterns were fed into three classification algorithms, Logistic Regression, Decision Tree and Random Forest. Out of these three methods, the Decision Tree classifier turned out to be the best in terms of recall which was the main used metric during the analysis.

All three models were able to identify clinically relevant patterns associated with being risk factors for childhood or late life obesity. The most frequently detected patterns mostly included parental BMI values, parental birth weight and gestational obesity showcasing the imprtance of these factors in the obesity of the infants. Although, alongside with clinically relevant patterns, some surprising and not expected variables and patterns turned out to be significant in the models. It is suspected that these variables were present due to inherent data unbalanedness.

Sequential pattern mining results were contrasted against previous itemset mining results, in order to understand the goodness of the method choice. This experiment showed that the results obtained with itemset mining on the same data performed dramatically better than the ones with sequential pattern mining.

The incorporation of increments were tested in a separate experimental case. The results indicated a slight decrease in performance.

In conclusion, the study was able to provide results that were mostly clinically relevant, and that can provide insights to obesity detection. With that being said, the used methods proven to be sub optimal as both the introduced data preprocessing and the pattern mining methods proved to be underperforming against other previously used methods.

9 Ethical social impact, sustainability and diversity

As a part of this given research, important ethical-social impact, sustainability, and diversity questions arose. This chapter aims to tackle the relevant questions and provide an outlook on such factors.

9.1 Ethical-social impact

The data is registered with study code 2010056 of the Clinical Research Ethics Committee of the Girona University Hospital Dr. Josep Trueta.

Privacy and confidentiality are key factors in biomedical research involving individuals. To ensure privacy, the original data was anonymized, making it impossible to trace back participants based on the final, utilized data. All participants gave informed consent to participate in the study.

To maintain maximum confidentiality, confidentiality agreements were signed by all agents performing the research and the research group providing the data. The data and results were not shared on any social platforms or within any online tool, ensuring the protection of sensitive information.

The project and the data collection were conducted in the cities of Girona and Figueres, Spain. No such study, related to obesity has conducted based on data from this region. This highlights the novelty provided by this subopulation.

9.2 Sustainability

The research prioritized models with low computational needs, thereby eliminating long computational times and excessive memory use. This approach not only enhanced efficiency but also reduced the environmental impact associated with high energy consumption.

The study was conducted with paperless documentation, effectively eliminating excessive paper use. Digital tools were employed for data collection, analysis, and reporting, aligning with sustainable practices and reducing the research's ecological footprint.

9.3 Diversity

Efforts were made to eliminate biases in the data collection process by ensuring a diverse and representative sample.

The collected data consisted and focused on fathers, mothers and their infants. This criteria excluded childless individuals. The distribution of the sexes of the infants were around the same, ensuring the exclusion of bias made by the gender of the infants.

Ethnicity-wise the study and the data handling were homogeneous. The inclusion criteria for the data included only Caucasian mothers. This was aimed to avoid genetic differences within the data, to minimize potenial additional confounding factors.

There were no exclusion criteria based on other factors such as income status, academic level of the parents or any other socio-economic status.

References

[1] WHO. *Obesity and overweight*. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (visited on 06/03/2024).

[2] Krushnapriya Sahoo et al. "Childhood obesity: causes and consequences". In: *Journal of Family Medicine and Primary Care* 4.2 (2015), pp. 187–192. ISSN: 2249-4863. DOI: 10.4103/2249-4863.154628.

[3] WHO. *Childhood obesity: five facts about the WHO European Region*. URL: https://www.who.int/europe/news/item/03-03-2023-childhood-obesity--five-facts-about-the-who-european-region (visited on 06/03/2024).

[4] John J. Reilly et al. "Early life risk factors for obesity in childhood: cohort study". In: *BMJ (Clinical research ed.)* 330.7504 (June 2005), p. 1357. ISSN: 1756-1833. DOI: 10.1136/bmj.38470.670903.E0.

[5] E. Oken, E. B. Levitan, and M. W. Gillman. "Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis". eng. In: *International Journal of Obesity (2005)* 32.2 (Feb. 2008), pp. 201–210. ISSN: 1476-5497. DOI: 10.1038/sj.ijo.0803760.

[6] Emily Oken et al. "Gestational weight gain and child adiposity at age 3 years". eng. In: *American Journal of Obstetrics and Gynecology* 196.4 (Apr. 2007), 322.e1–8. ISSN: 1097-6868. DOI: 10.1016/j.ajog.2006.11.027.

[7] David S. Ludwig and Janet Currie. "The association between pregnancy weight gain and birthweight: a within-family comparison". In: *Lancet* 376.9745 (2010), pp. 984–990. DOI: doi:10.1016/S0140-6736(10)60751-9.

[8] Renata Kuciene and Virginija Dulskiene. "Associations of maternal gestational hypertension with high blood pressure and overweight/obesity in their adolescent offspring: a retrospective cohort study". In: *Scientific Reports* 12 (Mar. 2022), p. 3800. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07903-z.

[9] Janis Baird et al. "Being big or growing fast: systematic review of size and growth in infancy and later obesity". In: *BMJ (Clinical research ed.)* 331.7522 (Oct. 2005), p. 929. ISSN: 1756-1833. DOI: 10.1136/bmj.38586.411273.E0.

[10] Salome Sunday and Zubair Kabir. "Impact of Carers' Smoking Status on Childhood Obesity in the Growing up in Ireland Cohort Study". In: *International Journal of Environmental Research and Public Health* 16.15 (Aug. 2019), p. 2759. ISSN: 1661-7827. DOI: 10.3390/ijerph16152759.

[11] Theresia M. Schnurr et al. “Smoking during pregnancy is associated with child overweight independent of maternal pre-pregnancy BMI and genetic predisposition to adiposity”. In: *Scientific Reports* 12.1 (Feb. 2022). Publisher: Nature Publishing Group, p. 3135. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07122-6.

[12] S. Arenz et al. “Breast-feeding and childhood obesity–a systematic review”. In: *Int J Obes Relat Metab Disord* 28.10 (2004), pp. 1247–1256. DOI: doi:10.1038/sj.ijo.0802758.

[13] Christopher G Owen et al. “Effect of infant feeding on the risk of obesity across the life course: a quantitative review of published evidence”. In: *Pediatrics* 115.5 (2005), pp. 1367–1377. URL: https://doi.org/10.1542/peds.2004-1176.

[14] Thomas Harder et al. “Duration of breastfeeding and risk of overweight: a meta-analysis”. In: *Am J Epidemiol* Sep 1;162.5 (2005), pp. 397–403. DOI: doi:10.1093/aje/kwi222..

[15] Elias Aguirre Rodríguez et al. “Machine learning Techniques to Predict Overweight or Obesity”. In: Valencia, Spain, Nov. 2021.

[16] Roupeng An, Jing Shen, and Yunyu Xiao. *Applications of Artificial Intelligence to Obesity Research: Scoping Review of Methodologies - PMC*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9856437/ (visited on 07/01/2024).

[17] Fati Musa, Frederick Basaky, and Osaghae E.O. “Obesity prediction using machine learning techniques”. In: *Journal of Applied Artificial Intelligence* 3 (June 2022), pp. 24–33. DOI: 10.48185/jaai.v3i1.470.

[18] Xueqin Pang et al. “Prediction of early childhood obesity with machine learning and electronic health record data”. In: *International Journal of Medical Informatics* 150, 104454 (2021). DOI: https://doi.org/10.1016/j.ijmedinf.2021.104454.

[19] Hagai Rossman et al. “Childhood obesity prediction from nationwide health records”. In: *medRxiv* (Jan. 2020), p. 2020.11.09.20228247. DOI: https://doi.org/10.1101/2020.11.09.20228247.

[20] Elizabeth A. Campbell et al. “Identification of temporal condition patterns associated with pediatric obesity incidence using sequence mining and big data”. In: *International Journal of Obesity* 44.8 (Aug. 2020). Publisher: Nature Publishing Group, pp. 1753–1765. ISSN: 1476-5497. DOI: 10.1038/s41366-020-0614-7.

[21] Chunlei Tang et al. “Medication Use for Childhood Pneumonia at a Children’s Hospital in Shanghai, China: Analysis of Pattern Mining Algorithms”. In: *JMIR Medical Informatics* 7.1 (Mar. 2019), e12577. DOI: 10.2196/12577.

[22] Elizabeth A Campbell, Ellen J Bass, and Aaron J Masino. “Temporal condition pattern mining in large, sparse electronic health record data: A case study in characterizing pediatric asthma”. In: *Journal of the American Medical Informatics Association : JAMIA* 27.4 (Feb. 2020), pp. 558–566. ISSN: 1067-5027. DOI: 10.1093/jamia/ocaa005.

[23] David Galera. “Longitudinal analysis of patterns and sequential rules with childhood obesity”. In: *MSc Thesis. University of Girona* (Sept. 2023).

[24] Marina Rodriguez Ros. “Application of sequential machine learning methods for the prediction of childhood obesity”. In: *Biomedical Engineering, Universitat de Girona,* (June 2023).

[25] Giuseppe Bonaccorso. *Machine learning algorithms: a reference guide to popular algorithms for data science and machine learning*. Birmingham Mumbai: Packt, 2017. ISBN: 978-1-78588-962-2.

[26] Philippe Fournier-Viger and Jerry Chun-Wei Lin. “A Survey of Sequential Pattern Mining”. In: *Data Science and Pattern Recognition (DSPR)* 1(1) (2017), pp. 55–77.

[27] Natalia Mordvanyuk, Albert Bifet, and Beatriz López. “VEPRECO: Vertical databases with pre-pruning strategies and common candidate selection policies to fasten sequential pattern mining”. In: *Expert Systems with Applications* 204 (Oct. 2022), p. 117517. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2022.117517.

[28] Philippe Fournier-Viger et al. “Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information”. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Vincent S. Tseng et al. Cham: Springer International Publishing, 2014, pp. 40–52. ISBN: 978-3-319-06608-0. DOI: 10.1007/978-3-319-06608-0_4.

[29] Jay Ayres et al. “Sequential PAttern mining using a bitmap representation”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD ’02. New York, NY, USA: Association for Computing Machinery, July 2002, pp. 429–435. ISBN: 978-1-58113-567-1. DOI: 10.1145/775047.775109.

[30] Mohammed J. Zaki. “SPADE: An Efficient Algorithm for Mining Frequent Sequences”. In: *Machine Learning* 42.1 (Jan. 2001), pp. 31–60. ISSN: 1573-0565. DOI: 10.1023/A:1007652502315.

[31] Jiawei Han et al. “FreeSpan: frequent pattern-projected sequential pattern mining”. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD ’00. New York, NY, USA: Association for Computing Machinery, Aug. 2000, pp. 355–359. ISBN: 978-1-58113-233-5. DOI: 10.1145/347090.347167.

[32] *SHAP documentation*. URL: https://shap.readthedocs.io/en/latest/ (visited on 07/24/2024).

[33] Hyo-Kyoung Nam and Kee-Hyoung Lee. “Small for gestational age and obesity: epidemiology and general risks”. In: *Annals of Pediatric Endocrinology & Metabolism* 23.1 (Mar. 2018), pp. 9–13. ISSN: 2287-1012. DOI: 10.6065/apem.2018.23.1.9.

[34] Yong Hee Hong and Ji-Eun Lee. “Large for Gestational Age and Obesity-Related Comorbidities”. In: *Journal of Obesity & Metabolic Syndrome* 30.2 (June 2021), pp. 124–131. ISSN: 2508-6235. DOI: 10.7570/jomes20130.

A Details of EDA

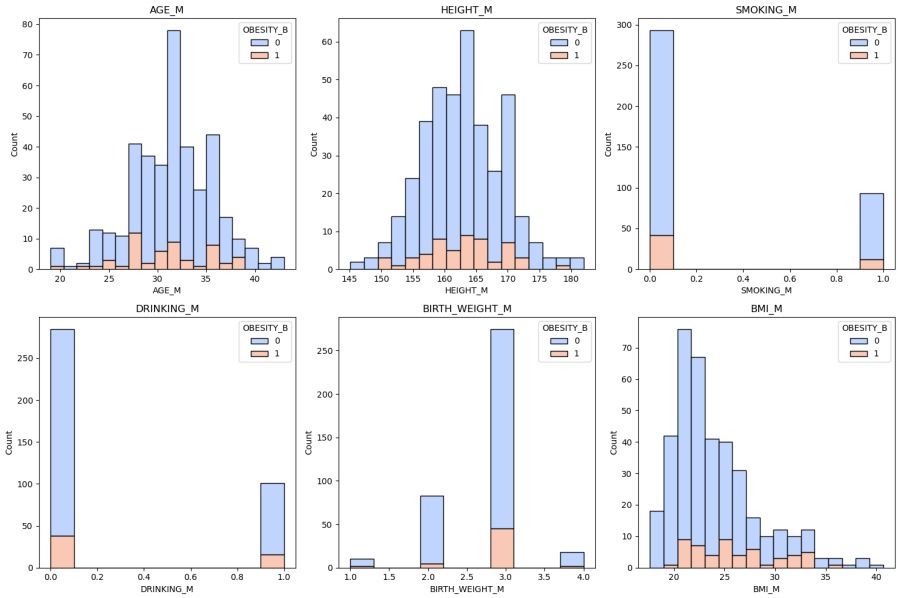

Figure 13: Distribution of the variables related to the mother - pre-pregnancy

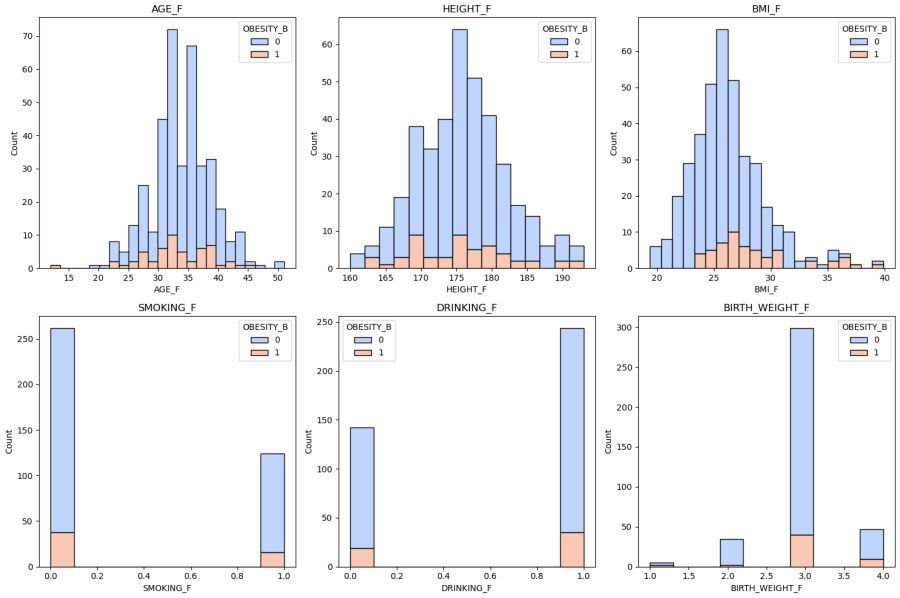

Figure 14: Distribution of the variables related to the father

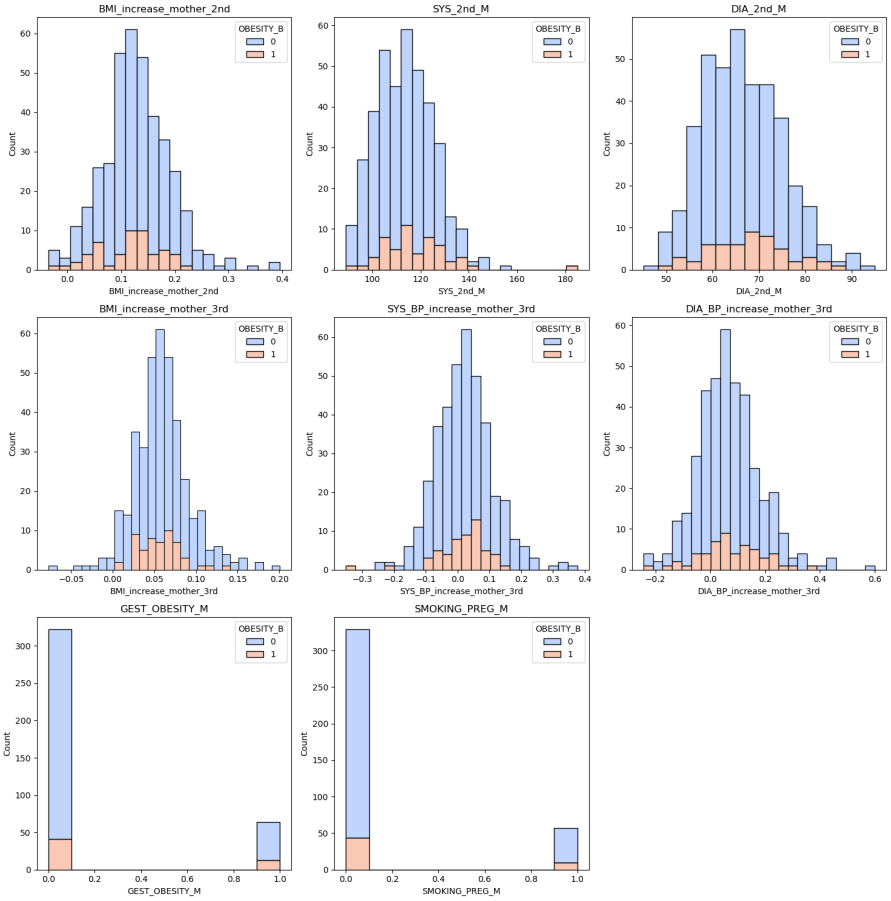

Figure 15: Distribution of the variables related to the mother - during pregnancy (2nd and 3rd trimester)

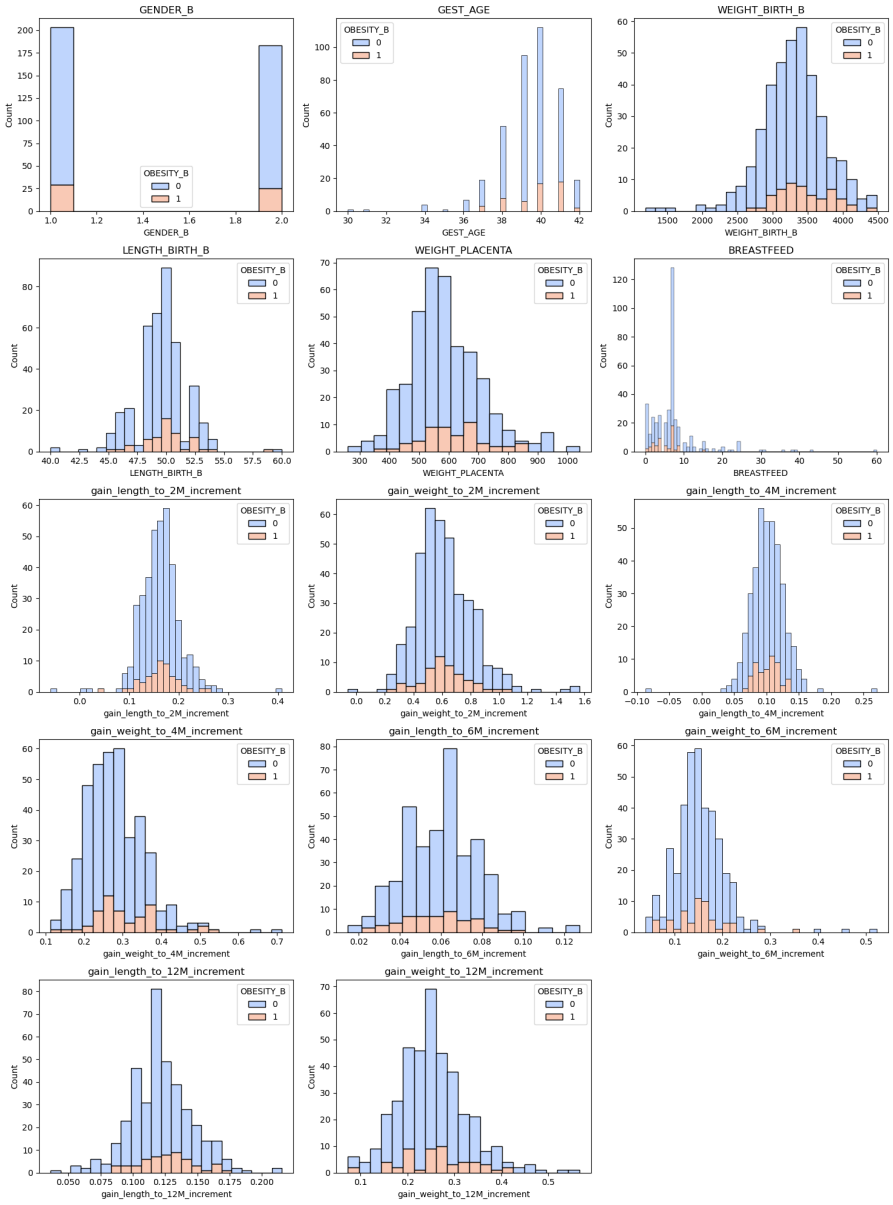

Figure 16: Distribution of the variables related to the infant

B Codification of the discretized data

Code	Description
1	AGE_M $\leq$ 29
2	29 $<$ AGE_M $\leq$ 31.5
3	31.5 $<$ AGE_M $\leq$ 34
4	34 $<$ AGE_M
11	HEIGHT_M $\leq$ 159
12	159 $<$ HEIGHT_M $\leq$ 163
13	163 $<$ HEIGHT_M $\leq$ 167
14	167 $<$ HEIGHT_M
21	SMOKING_M $==$ 0
22	SMOKING_M $==$ 1
31	DRINKING_M $==$ 0
32	DRINKING_M $==$ 1
41	BIRTH_WEIGHT_M $==$ 1
42	BIRTH_WEIGHT_M $==$ 2
43	BIRTH_WEIGHT_M $==$ 3
44	BIRTH_WEIGHT_M $==$ 4
51	IS_FIRST_CHILD $==$ 1
52	IS_FIRST_CHILD $==$ 2
61	BMI_M $<$ 18.5
62	18.5 $\leq$ BMI_M $<$ 25
63	25 $\leq$ BMI_M $<$ 30
64	30 $\leq$ BMI_M $<$ 35
65	35 $\leq$ BMI_M
71	AGE_F $\leq$ 31
72	31 $<$ AGE_F $\leq$ 34
73	34 $<$ AGE_F $\leq$ 47
74	47 $\leq$ AGE_F
81	HEIGHT_F $\leq$ 172
82	172 $<$ HEIGHT_F $\leq$ 176
83	176 $<$ HEIGHT_F $\leq$ 180
84	180 $<$ HEIGHT_F
91	BMI_F $<$ 18.5
92	18.5 $\leq$ BMI_F $<$ 25
93	25 $\leq$ BMI_F $<$ 30
94	30 $\leq$ BMI_F $\leq$ 35
95	35 $\leq$ BMI_F

Code	Description
101	SMOKING_F == 0
102	SMOKING_F == 1
111	DRINKING_F == 0
112	DRINKING_F == 1
121	BIRTH_WEIGHT_F == 1
122	BIRTH_WEIGHT_F == 2
123	BIRTH_WEIGHT_F == 3
124	BIRTH_WEIGHT_F == 4
131	BMI_increase_mother_2nd < 0.5
132	0.5 ≤ BMI_increase_mother_2nd < 1.5
133	1.5 ≤ BMI_increase_mother_2nd < 2.5
134	2.5 ≤ BMI_increase_mother_2nd
141	SYS_2nd_M < 115
142	115 ≤ SYS_2nd_M < 120
143	120 ≤ SYS_2nd_M < 139
144	139 ≤ SYS_2nd_M
151	DIA_2nd_M < 60
152	60 ≤ DIA_2nd_M < 70
153	70 ≤ DIA_2nd_M < 80
154	80 ≤ DIA_2nd_M
161	BMI_increase_mother_3rd < 0
162	0 ≤ BMI_increase_mother_3rd < 0.05
163	0.05 ≤ BMI_increase_mother_3rd < 0.1
164	0.1 ≤ BMI_increase_mother_3rd < 0.15
165	0.15 ≤ BMI_increase_mother_3rd
171	SYS_BP_increase_mother_3rd < -0.4
172	-0.4 ≤ SYS_BP_increase_mother_3rd < 0.018
173	0.018 ≤ SYS_BP_increase_mother_3rd < 0.072
174	0.072 ≤ SYS_BP_increase_mother_3rd
181	DIA_BP_increase_mother_3rd < 0
182	0 ≤ DIA_BP_increase_mother_3rd < 0.059
183	0.059 ≤ DIA_BP_increase_mother_3rd < 0.129
184	0.129 ≤ DIA_BP_increase_mother_3rd
191	GEST_OBESITY_M == 0
192	GEST_OBESITY_M == 1
201	SMOKING_PREG_M == 0
202	SMOKING_PREG_M == 1
211	GENDER_B == 1

Code	Description
212	GENDER_B == 0
221	GEST_AGE ≤ 34
222	34 < GEST_AGE ≤ 36
223	36 < GEST_AGE ≤ 38
224	38 ≤ GEST_AGE
231	WEIGHT_BIRTH_B ≤ 3000
232	3000 < WEIGHT_BIRTH_B ≤ 3500
233	3500 < WEIGHT_BIRTH_B ≤ 4000
234	4000 ≤ WEIGHT_BIRTH_B
241	LENGTH_BIRTH_B < 48
242	48 ≤ LENGTH_BIRTH_B < 50
243	50 ≤ LENGTH_BIRTH_B < 51
244	51 ≤ LENGTH_BIRTH_B
251	WEIGHT_PLACENTA < 500
252	500 ≤ WEIGHT_PLACENTA < 580
253	580 ≤ WEIGHT_PLACENTA < 660
254	660 ≤ WEIGHT_PLACENTA
261	BREASTFEED < 4
262	4 ≤ BREASTFEED < 7
263	BREASTFEED == 7
264	7 < BREASTFEED
271	gain_length_to_2M_increment < 0.125
272	0.125 ≤ gain_length_to_2M_increment < 0.175
273	0.175 ≤ gain_length_to_2M_increment < 0.225
274	0.225 ≤ gain_length_to_2M_increment
281	gain_weight_to_2M_increment < 0.4
282	0.4 ≤ gain_weight_to_2M_increment < 0.6
283	0.6 ≤ gain_weight_to_2M_increment < 0.8
284	0.8 ≤ gain_weight_to_2M_increment
291	gain_length_to_4M_increment < 0.075
292	0.075 ≤ gain_length_to_4M_increment < 0.1
293	0.1 ≤ gain_length_to_4M_increment < 0.125
294	0.125 ≤ gain_length_to_4M_increment
301	gain_weight_to_4M_increment < 0.2
302	0.2 ≤ gain_weight_to_4M_increment ≤ 0.275
303	0.275 < gain_weight_to_4M_increment < 0.35
304	0.35 ≤ gain_weight_to_4M_increment
311	gain_length_to_6M_increment < 0.04

Code	Description
312	$0.04 \leq$ gain_length_to_6M_increment < 0.06
313	$0.06 \leq$ gain_length_to_6M_increment < 0.08
314	$0.08 \leq$ gain_length_to_6M_increment
321	gain_weight_to_6M_increment < 0.1
322	$0.1 \leq$ gain_weight_to_6M_increment < 0.15
323	$0.15 \leq$ gain_weight_to_6M_increment < 0.2
324	$0.2 \leq$ gain_weight_to_6M_increment
331	gain_length_to_12M_increment < 0.1
332	$0.1 \leq$ gain_length_to_12M_increment < 0.125
333	$0.125 \leq$ gain_length_to_12M_increment < 0.15
334	$0.15 \leq$ gain_length_to_12M_increment
341	gain_weight_to_12M_increment < 0.2
342	$0.2 \leq$ gain_weight_to_12M_increment < 0.25
343	$0.25 \leq$ gain_weight_to_12M_increment < 0.3
344	$0.3 \leq$ gain_weight_to_12M_increment

Table 6: Codification of the discretized data