Instituto Superior de Engenharia
Politécnico de Coimbra

DEPARTMENT OF SYSTEMS AND COMPUTER ENGINEERING

AI Models for Obesity Risk Assessment

Project Report to fulfill the Master's degree in Informatics Engineering

Specialization in Intelligent Data Analysis

Author

Sandra Nathaly Aguilar Alvarez

Supervisor

PhD. Francisco Pereira

Co-Supervisor

PhD. António Cruz

Coimbra, October 2025

INSTITUTO POLITÉCNICO DE COIMBRA

INSTITUTO SUPERIOR DE ENGENHARIA DE COIMBRA

Abstract

Childhood overweight and obesity reflect interconnected behavioural, environmental and social influences. Within the Horizon-Europe *PAS GRAS* initiative and its longitudinal *GicO* study, this work investigates whether data-driven methods can reveal meaningful patterns that support detection and prevention at community level. The main objective is to develop Artificial Intelligent models that identify determinants of childhood obesity risk by integrating multi-source information into profiles and indicators that are easily interpretable for early action in real-world settings.

Using harmonised datasets covering children's body composition and physical fitness together with family-context measures, we apply unsupervised learning to characterise population structure and relate profiles to broader behaviours and context. A brief longitudinal view illustrates how children weight status indicators evolve across assessment moments.

Results reveal three clear, interpretable clusters spanning healthier, intermediate and higher-risk profiles. The principal contributions are: (i) an integrated, transparent pipeline from data cleaning to modelling; (ii) visual profile summaries that aid communication with practitioners; and (iii) concise, actionable indicators that connect model outputs to prevention and screening. The materials aim to inform multidisciplinary teams working in community programmes and to guide tailored recommendations for specific groups.

Keywords: Artificial Intelligence, childhood obesity, data integration, high-dimensional reduction, clustering, profiling, SHAP explainability, BMI percentile, longitudinal analysis.

Resumo

O excesso de peso e a obesidade infantil refletem influências interligadas de natureza comportamental, ambiental e social. No âmbito da iniciativa europeia *Horizon-Europe PAS GRAS* e do seu estudo longitudinal *GicO*, este trabalho investiga se métodos baseados em dados podem revelar padrões significativos que apoiem a deteção e prevenção a nível comunitário. O principal objetivo é desenvolver modelos de Inteligência Artificial capazes de identificar determinantes do risco de obesidade infantil, integrando informação de múltiplas fontes em perfis e indicadores facilmente interpretáveis, de modo a permitir uma ação precoce em contextos reais.

Recorrendo a conjuntos de dados harmonizados que abrangem a composição corporal e a aptidão física das crianças, em conjunto com medidas do contexto familiar, aplicam-se técnicas de aprendizagem não supervisionada para caracterizar a estrutura populacional e relacionar os perfis com comportamentos e contextos mais amplos. Uma breve análise longitudinal ilustra como os indicadores do estado ponderal das crianças evoluem ao longo dos momentos de avaliação.

Os resultados revelam três clusters claros e interpretáveis, que abrangem perfis mais saudáveis, intermédios e de maior risco. As principais contribuições são: (i) um pipeline integrado e transparente desde a limpeza dos dados até à modelação; (ii) resumos visuais dos perfis que facilitam a comunicação com profissionais de saúde e educação; e (iii) indicadores concisos e acionáveis que ligam os resultados dos modelos à prevenção e rastreio. Os materiais desenvolvidos visam apoiar equipas multidisciplinares envolvidas em programas comunitários e orientar recomendações adaptadas a cada caso.

Palavras-chave: Inteligência Artificial, obesidade infantil, integração de dados, redução de dimensionalidade, clustering, caracterização de perfis, explicabilidade SHAP, percentil de IMC, análise longitudinal.

Epigraph

Data science isn't about the quantity of data but rather the quality.
Joo Ann Lee

Dedication

To God, the source and sustainer of every good thing in my life, who has guided each step of this journey.

To Ricardo, my boyfriend and best friend, your unwavering support and belief are behind every page of this work.

To my parents, with profound gratitude for your love, sacrifices, and constant encouragement.

Acknowledgments

I am profoundly grateful to Portugal for opening its doors to me. This country has enabled me to study, to work, and to live with a high quality of life, and I hold that privilege with enduring gratitude.

My sincere thanks go to the Coimbra Institute of Engineering (ISEC), which has become my second *alma mater*. The training, intellectual community, and supportive environment at ISEC have shaped both the rigour and the curiosity reflected in this work.

I wish to express my heartfelt gratitude to my supervisor, Prof. PhD. Francisco Pereira, for his guidance, high standards, and steady encouragement. I am equally grateful to my co-supervisor, PhD. António Cruz, for his insightful feedback, clarity, and generosity with his time.

I also thank the team members I have worked with: Andreia Silva, Marisa Azul, Paulo Matafome, and Prof. Luís Rama, whose expertise and support were invaluable to the progress of this research.

To everyone who contributed, colleagues: Lourenço Carlos, Vasco Oliveira and Rui Oliveira, friends and family, thank you for your kind words and practical help along the way.

Table of contents

List of tables

List of figures

List of abbreviations

AI	Artificial Intelligence
BIA	Bioelectrical Impedance Analysis
CMJ	Countermovement Jump
CNC–UC	Centre for Neuroscience and Cell Biology, University of Coimbra
CRMU	Cardiorespiratory and Metabolic Unit
DASS	Depression Anxiety Stress Scales
DQI	Diet Quality Index
EDA	Exploratory Data Analysis
FIS	Food Insecurity Score
GAI	Green Access Index
GicO	Gymnastics Against Obesity Study
HEI	Healthy Eating Index–2010
ID	Identifier
IDE	Integrated Development Environment
ISEC	Instituto Superior de Engenharia de Coimbra
KDE	Kernel Density Estimate
LASSO	Least Absolute Shrinkage and Selection Operator
MDS	Mediterranean Diet Score
MDS (stats)	Multidimensional Scaling
ML	Machine Learning
OECD	Organisation for Economic Co–operation and Development
PAI	Physical Activity Index
PCA	Principal Component Analysis
PSQI	Pittsburgh Sleep Quality Index
SAD	Stress, Anxiety, and Depression
SES	Socio–Economic Status
SHAP	SHapley Additive exPlanations
SJ	Squat Jump
SQI	Sleep Quality Index
STI	Screen Time–Sedentary Index
t–SNE	t–Distributed Stochastic Neighbour Embedding
WHO	World Health Organization

List of Symbols

α	Lasso regularisation strength (penalty weight).
BCM	Body Cell Mass
BMI	Body Mass Index
BMR	Basal Metabolic Rate
cm, m, kg	Centimetre
FFM, FM	Fat–Free Mass, Fat Mass
FFMI	Fat–Free Mass Index
h, min, s	Hour, minute, second
h/week	Hours per week
IQR	Interquartile range $(Q_3 - Q_1)$
kg	Kilogram
$\mathrm{kg/m^2}$	Kilograms per square metre (BMI unit)
kHz	Kilohertz
L_v	Lower bound of the device-provided Normal Range for variable v (observation-specific).
L_1	L1 penalty.
MET	Metabolic Equivalent of Task
m	Metre
mL	Millilitre
$\mathrm{mL \cdot kg^{-1} \cdot min^{-1}}$	Unit for $\dot{V}O_{2\,\mathrm{max}}$
n	Sample size
p	p-value
Q_1, Q_3	First and third quartiles
SD	Standard Deviation
SLM	Soft Lean Mass
SMM	Skeletal Muscle Mass
TBW, ICW, ECW	Total, intra- and extracellular body water
U_v	Upper bound of the device-provided Normal Range for variable v (observation-specific).
v	Generic continuous variable
$VO_{2\,\mathrm{max}}$	Maximal oxygen uptake
W	Watt

1 Introduction

Childhood overweight and obesity are complex, multi-factorial conditions that emerge from linked behavioural, environmental, and socio-economic determinants. Despite extensive evidence on risk factors, Europe has not yet succeeded in stopping the upward trend. Current projections indicate that, unless effective actions are scaled, the pressure on the health system from obesity will rise substantially over the next decade. These challenges call for approaches that integrate data across domains and turn them into practical evidence for screening and early prevention.

Within this broader landscape, the *Gymnastics Against Obesity* (GicO) study provides a uniquely rich setting. The study is embedded within the European research project **PAS GRAS**[1], *Metabolic Risk Reduction, Environmental and Behavioural Determinants of Obesity in Children, Adolescents and Young Adults* [1]. PAS GRAS is a five-year (2023–2028) project funded by the European Union's Horizon Europe Programme; it seeks to prevent and reverse overweight and obesity by integrating biomedical, behavioural, and environmental perspectives to deliver personalised early-risk assessment tools and targeted, community-ready interventions for **children, adolescents, and young people**. It is carried out by a *multidisciplinary team*, bringing together expertise from biomedical sciences, public health, behavioural sciences, environmental epidemiology, and data science. In this context, the GicO study, led by the University of Coimbra[2], is a programme that collects multi-source, longitudinal data to support early screening and prevention. The GicO study will be presented in greater detail in Chapter 2.

This work leverages the GicO/PAS GRAS data environment to investigate whether artificial intelligence (AI) can uncover child profiles linked to obesity risk. It also examines how those profiles relate to family behaviours and contexts. The aim is to identify key determinants of obesity risk in the GicO sample and to generate evidence that helps the multidisciplinary team tailor recommendations to specific groups. This project is carried out under the *Master in Informatics Engineering, Specialisation in Intelligent Data Analysis* at the Coimbra Institute of Engineering (ISEC).

[1] `https://pasgras.eu/en`
[2] `https://www.uc.pt/`

1.1 Objectives

This work aims to develop AI models to identify determinants of childhood obesity risk, integrating the multi-source GicO datasets, translating evidence into profiles and indicators that support early screening and prevention in community settings.

The project includes integration within the PAS GRAS research team, collaborating closely on data analysis tasks and aligning methods, documentation, and results under the guidance of the team's experts and their domain knowledge.

It further comprises the pre-processing and exploratory data analysis (EDA) of datasets collected by the multidisciplinary team, harmonisation across sources, and transparent descriptive insights.

We analyse two complementary sources: (i) questionnaires completed by the parents of the participating children, capturing family behaviour and context; and (ii) child data comprising body composition, cardiorespiratory and functional fitness, and qualitative evidence from children's drawings that reflect perceptions of food, activity, and everyday environment.

Building on this foundation, the work designs and validates *unsupervised* models to uncover patterns associated with the underlying drivers of obesity among children participating in the GicO study.

1.2 Main contributions

This project makes the following contributions:

- A preprocessing and data–engineering framework that harmonises heterogeneous, multi-source (multimodal), high-dimensional datasets from the GicO/PAS GRAS context.
- A thorough exploratory data analysis (EDA) that characterises data distributions, patterns, and cross-domain associations.
- An integrated, multi-source risk-profiling pipeline for childhood obesity that jointly leverages information from children and their parents.
- Modelling workflows that profile groups and identify those with elevated overweight/obesity risk, combining unsupervised discovery with predictive modelling.
- A model-interpretation layer based on SHAP (*SHapley Additive exPlanations*) to support transparent and expert-guided insights.

1.3 Task management

This project was conducted over approximately one scholar year (September 2024–July 2025) and was organised into five workstreams, executed iteratively with partial overlap. A detailed work plan is provided in Appendix A.

- **T1 — Undestanding the problem**: literature review and scoping of the problem domain.
- **T2 — Exploratory analysis**: data understanding, pre-processing, and exploratory data analysis (EDA).
- **T3 — Unsupervised analysis**: design and validation of unsupervised models.
- **T4 — Profiling groups**: interpretation of discovered clusters, construction of risk profiles, alignment with expert knowledge, and derivation of actionable indicators for early screening.
- **T5 — Documentation**: production of reproducible artefacts and thesis writing.

Figure 1.1 shows the time distribution of each task across the project period.

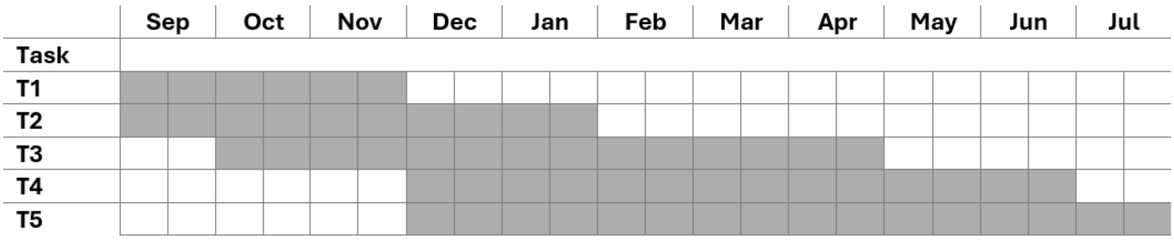


Figure 1.1: Time distribution of tasks (Sep 2024–Sep 2025).

1.4 Document structure

This document is structured into 7 different chapters: Chapter 2 presents the GicO study, its datasets and baseline characteristics. Chapter 3 reports data cleaning, preprocessing and feature engineering, including the construction of questionnaire-derived indices. Chapter 4 details the clustering methodology, profiling strategy, cross-domain association analyses, interpretability, and the longitudinal component. Chapter 5 presents the empirical results for cluster identification and profiling, associations with other domains, SHAP-based insights, and BMI-percentile trajectories. Chapter 6 discusses the findings against prior evidence, implications for screening, strengths/limitations, and robustness checks. Chapter 7 concludes and outlines future work.

2 The Gymnastics Against Obesity Study (GicO)

This chapter frames the data and study context of the project. It first situates the work within the PAS GRAS programme and outlines the scope and aims of the Gymnastics Against Obesity (GicO) study. We then introduce the multidisciplinary team and provide an overview of the datasets to them. Next, we provide concise descriptions of each dataset: the parental questionnaire (completed by parents) and, for children, the body composition measurements, cardiorespiratory assessments, functional performance tests, and drawings. The chapter closes with a brief overview of related work that motivates the modelling strategy used later in the document.

Specifically, the GicO [1] Study [2] is a longitudinal, interdisciplinary research initiative designed to assess the effectiveness of integrated interventions combining physical exercise, healthy eating, and science and health literacy in preventing and treating overweight and obesity among school-aged children and their families. The study is embedded within the broader European research project **PAS GRAS** mentioned in the previous chapter.

The GicO Study adopts a participatory approach by actively involving children, families, and a multidisciplinary team of researchers from sports sciences, biomedicine, nutrition, ecology, geography, health communication, and visual arts. This approach aims to generate evidence-based strategies for obesity prevention that can be adapted to other European contexts. It also includes practical assessments of children's physical abilities, conducted in a controlled indoor environment, as illustrated in Figure 2.1.

The PAS GRAS project addresses obesity as a multifactorial public health challenge, integrating biomedical, behavioural, environmental, and socioeconomic perspectives. In this framework, GicO operates as a local case study in Coimbra, Portugal, with a focus on:

- Early prevention during critical developmental stages (ages 6–8).
- Family-based interventions to ensure sustained lifestyle changes.
- Multi-method data collection, combining physiological measures, environmental assessments, and qualitative insights.

The overall purpose of the GicO study is to evaluate the impact of a gymnastics-based intervention, complemented by nutritional education and health literacy activities, on preventing and reducing overweight and obesity in children and their families. The

[1] Original study name in Portuguese: *Ginástica Contra a Obesidade* [2].

Figure 2.1: Functional performance assessment during the GicO Study [3].

study focuses on monitoring changes in physical fitness, body composition, behaviours, and perceptions related to diet, activity, sleep, and wellbeing, while exploring how these are influenced by socio-demographic and environmental factors. It also seeks to engage families through participatory activities and to develop educational resources that promote sustainable, healthy lifestyles and can be adapted to other contexts.

2.1 Multidisciplinary team

GicO is led by the University of Coimbra within the PAS GRAS framework and brings together a *multidisciplinary* team spanning sports and exercise sciences, nutrition, psychology, epidemiology, and education sciences, among others. The domain experts who collaborated most closely with this work are affiliated with the Faculty of Sport Sciences and Physical Education of the University of Coimbra (FCDEF-UC) [4] and the Centre for Neuroscience and Cell Biology of the University of Coimbra (CNC-UC) [5]:

- **Andreia Filipa Neves da Silva**, PhD — Researcher at CNC-UC.
- **Luís Manuel Pinto Lopes Rama**, PhD — Professor at FCDEF-UC.
- **Anabela Marisa Azul**, PhD — Researcher at CNC-UC.
- **Paulo Nuno Centeio Matafome**, PhD — Researcher at CNC-UC.

Their expertise informed study design and protocol calibration, variable curation and harmonisation, and the interpretation of the exploratory and modelling results throughout this work.

2.2 Datasets Overview

The GicO Study follows a longitudinal design with three measurement phases conducted during 2024. These phases are referred to as Time Points: T1, T2, and T3, and they may allow us to observe how children's characteristics evolve over time. The interval between time points is approximately six months. The data collection protocol produced five complementary datasets that together cover behavioural context at home and children's physiological status and perceptions.

Figure 2.2 presents a matrix-style view in which the rows are the five datasets (parental questionnaire, body composition, cardiorespiratory fitness, functional fitness, and drawings) and the columns are the three time points (T1–T3). Markers indicate who contributed data, green circles for children and an orange square for parents, with the participant count n shown beside each marker. T1 is highlighted as the baseline and is the most comprehensive collection point because it includes all five datasets; it is therefore the principal reference on which the subsequent data analysis is based.

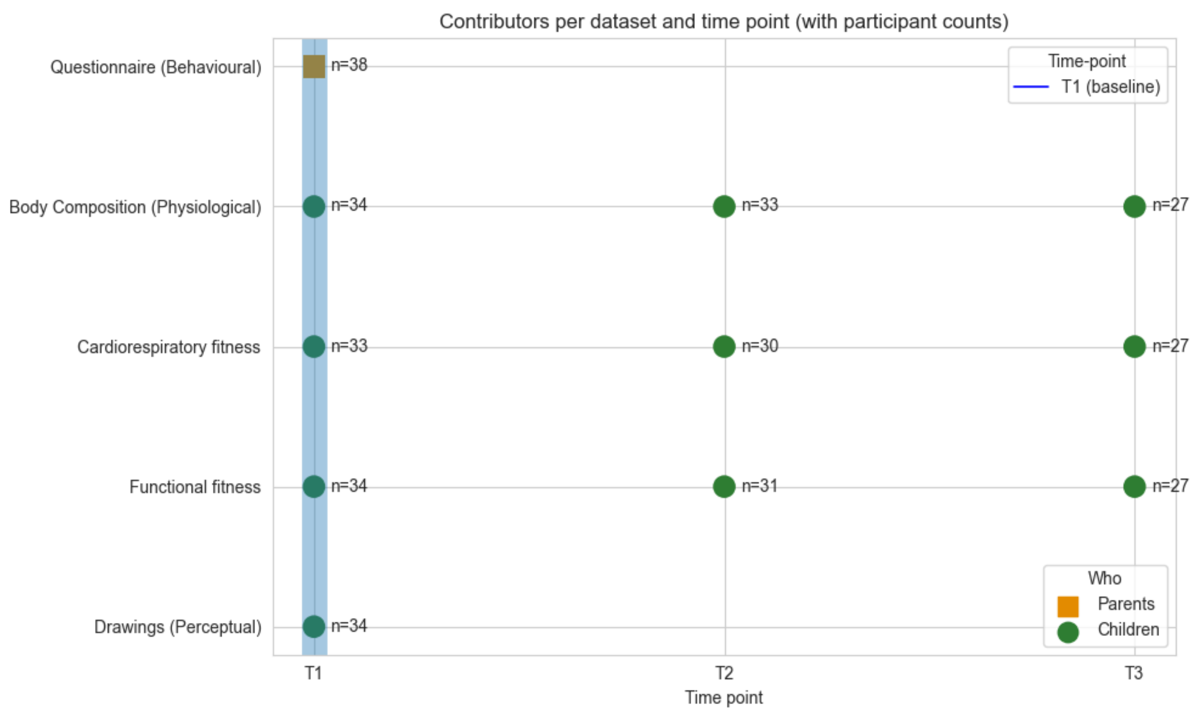


Figure 2.2: Who contributes to each dataset across time points.

The five datasets are interrelated through a shared identifier system. The *Body Composition*, *Cardiorespiratory*, *Functional*, and *Drawings* datasets each store a unique child identifier (e.g., `N001`, `N002`), which acts as the primary key to link records across these sources. The *Parental Questionnaire* uses extended identifiers such as `N001A` or `N001B`, where the prefix matches the child's ID and the suffix denotes the responding parent.

1. **Parental questionnaire (T1 only, parents):** Applied at the start of the study (T1), this survey captures household demographics, socio-economic context, health

and literacy, food practices, physical-activity routines, screen-time rules, sleep habits, and parents' perceptions and concerns about their wellbeing. Its objective is to characterise the family environment and behaviours that may shape obesity risk.

2. **Body composition (T1–T3, children):** Measurements such as BMI and bio impedance derived indices (e.g., fat mass, fat-free mass, water) quantify children's physiological status. The objective is to track growth and body composition changes.

3. **Functional performance (T1–T3, children):** Field tests (e.g., handgrip, squat jump, 20-metre sprint, sit-and-reach) assess neuromuscular fitness and motor competence. The objective is to understand how functional capacity relates to overall health.

4. **Cardiorespiratory fitness (T1–T3, children):** Aerobic capacity is evaluated through standardised field assessments and derived indicators (e.g., estimated VO_2-related metrics).

5. **Drawings (T1 only, children):** A single drawing dataset collected at T1 captures children's perceptions around food, activity, and what most catches their eye day to day. The objective is to complement measured behaviours with perceptual cues that may signal preferences or awareness.

2.3 Sample Characteristics at Baseline (T1): Parents and Children

Here we give a clear snapshot of who took part at baseline (T1). First, we describe the parents: their age, gender, education, health, and daily habits. Then, we summarise the children's measurements (age, heigh,t and weight).

2.3.1 Parents

A total of 38 parents completed the online questionnaire. Their ages ranged from 34 to 55 years, with a mean of 43.9 years (SD = 4.1). Quartile values (Q1 = 41.3; median = 43.5; Q3 = 46.0) indicate that most parents were in their early to mid-forties; the full age distribution, including Q1, median, Q3, whiskers (min–max) and a ±SD is shown in Figure 2.3. The gender distribution was relatively balanced, with 22 women (57.9%) and 16 men (42.1%) (Figure 2.4). Regarding educational qualifications, a large share held higher education degrees, 36.8% bachelor's, 31.6% master's, and 18.4% PhD, while 13.1% had completed secondary school or less (Figure 2.5).

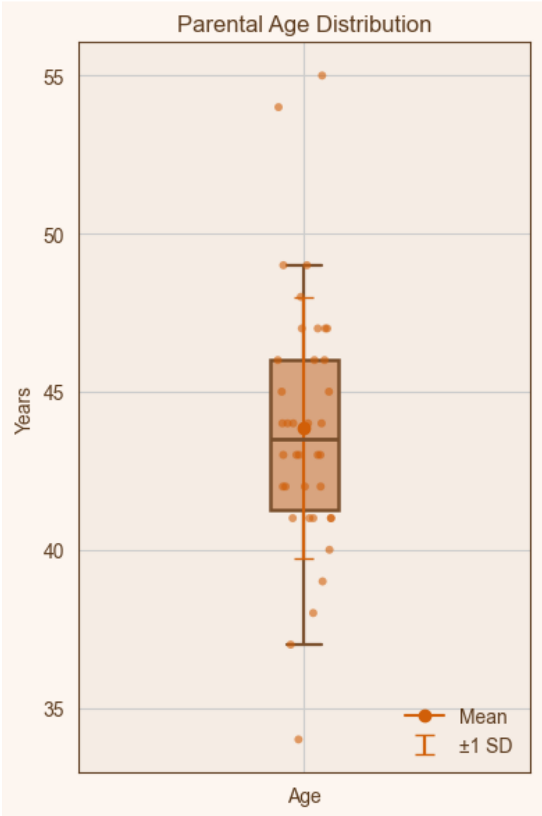


Figure 2.3: Parental age distribution (T1).

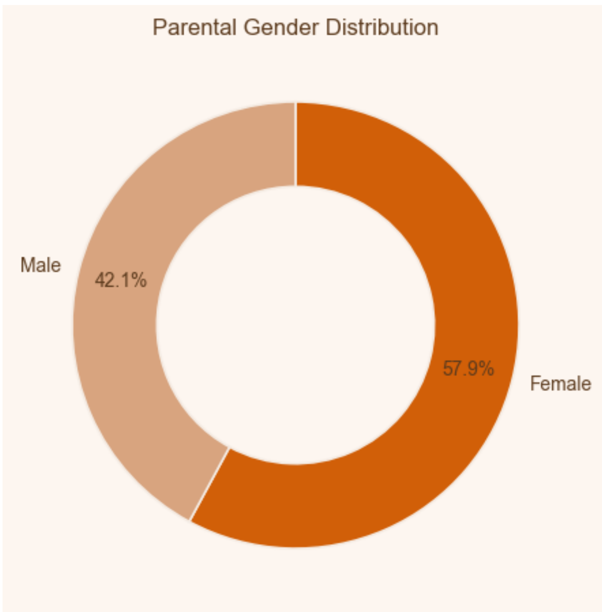


Figure 2.4: Parental gender distribution (T1).

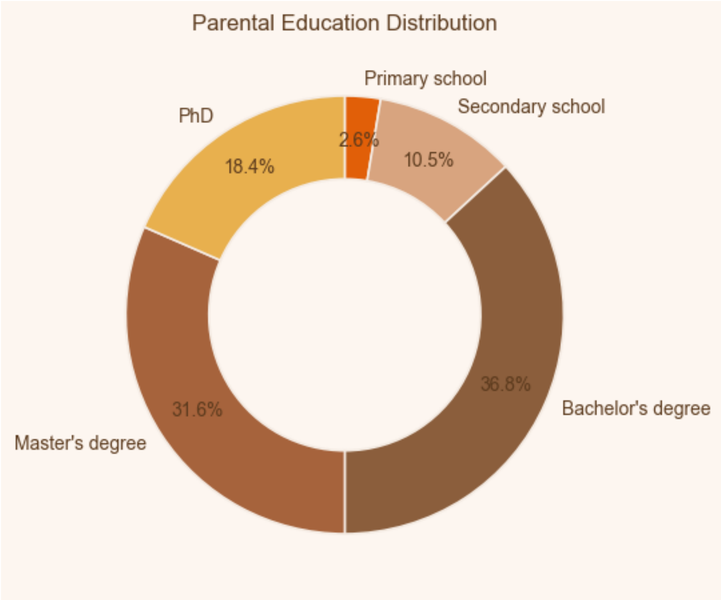


Figure 2.5: Parental education distribution (T1).

Lifestyle and environmental factors

In terms of weight status, 45% of the parents were classified as having a normal weight, while 37% were overweight, 13% obese, and 5% underweight (Table 2.1). Notably, 26% reported having high cholesterol and 13% reported suffering from anxiety.

Concerning household routines, 60.5% of parents stated they always cooked meals at home, 23.7% cooked occasionally, and 15.8% reported never cooking. Sleep duration varied: 63% of participants slept between 5 and 7 hours per night, 34% slept more than 7 hours, and only 3% slept fewer than 5 hours.

Physical activity patterns also showed diversity. Half of the parents (50%) did not engage in any regular exercise. Among the rest, 29% reported doing between 0.75 and 3 hours of physical activity per week, while 21% exercised for more than 3 hours weekly.

2.3.2 Children

At baseline (T1), 34 children completed the body composition, functional, and cardiorespiratory assessments, as well as providing drawings. The average age was 6.7 years (SD = 0.3), ranging from 6.2 to 7.2 years. Quartiles confirm the narrow distribution within this school-age range (Q1 = 6.3; median = 6.8; Q3 = 6.9), as illustrated in Figure 2.6. Girls represented 76% of the sample, while boys accounted for 24% (Figure 2.7).

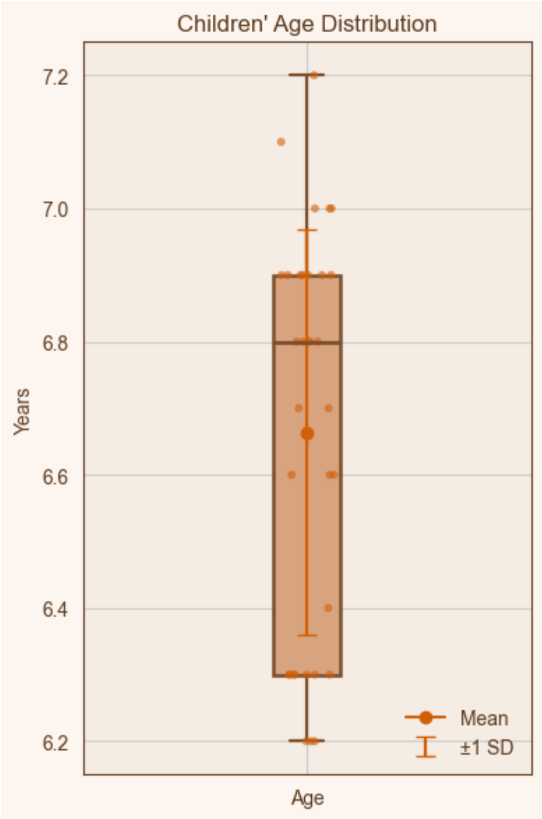


Figure 2.6: Children's age distribution (T1). Box shows Q1–Q3 with median; whiskers denote min–max; point and bar indicate mean and mean ±SD.

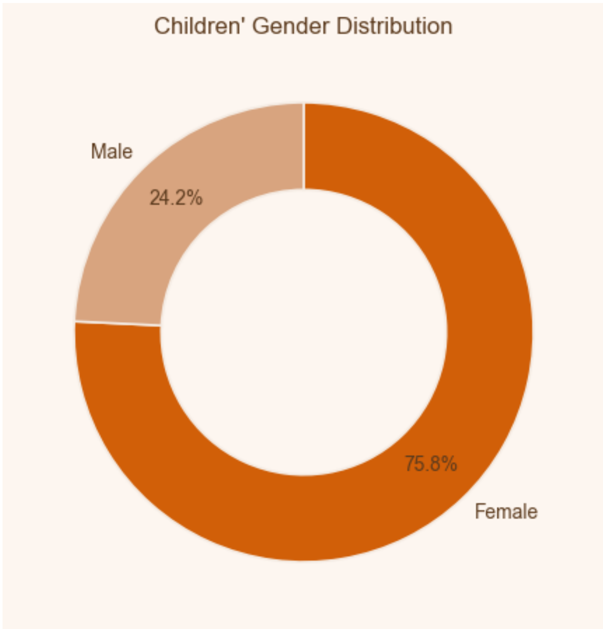


Figure 2.7: Children's gender distribution (T1).

Table 2.1: Health characteristics of participating parents.

Characteristic	Number (%)
Total parents	38 (100)
Weight status	
Underweight	2 (5)
Normal weight	17 (45)
Overweight	14 (37)
Obese	5 (13)
Health conditions	
High cholesterol	10 (26)
Anxiety	5 (13)
Cooking habit	
Never	6 (15.8)
Sometimes	9 (23.7)
Always	23 (60.5)
Sleep duration	
< 5 hours	1 (3)
5–7 hours	24 (63)
> 7 hours	13 (34)
Exercise per week	
0 hours	19 (50)
0.75–3 hours	11 (29)
> 3 hours	8 (21)

Anthropometric information

Regarding anthropometric characteristics, most children fell within the 120–125 cm height range, accounting for 35.3% of the sample, followed by 26.5% in the 115-120 cm range and 20.6% in the 110–115 cm range. Taller children were less common, with only 2.9% each in the 130–135 cm and 135–140 cm intervals.

In terms of body weight, half of the children weighed between 20 and 25 kg, while 29.4% were in the 25–30 kg range. A smaller proportion, 17.6%, weighed less than 20 kg, and only one child, representing 2.9% of the sample, was in the 35-40 kg range.

A detailed summary of this anthropometric information is presented in Table 2.2.

2.4 Description of Each Dataset

Having framed the GicO project and its inherently multi-dimensional, multi-source and longitudinal design, covering parental questionnaires, body composition, cardiorespiratory and functional assessments, and children's drawings, and having provided a high-level overview of each dataset's scope together with the characterisation of parents and their children, this section now moves from narrative context to technical detail. We set out each dataset's data structure, column definitions, and data types.

Table 2.2: Anthropometric information of children (T1).

Characteristic	**Number (%)**
Total children	34 (100)
Height (cm)	
110–115	7 (20.6)
115–120	9 (26.5)
120–125	12 (35.3)
125–130	4 (11.8)
130–135	1 (2.9)
135–140	1 (2.9)
Weight (kg)	
< 20	6 (17.6)
20–25	17 (50.0)
25–30	10 (29.4)
35–40	1 (2.9)

2.4.1 Parental questionnaire data

The questionnaire administered to parents comprised 89 items and was designed to capture detailed information about household routines, family lifestyle, and the child's social, digital, and nutritional environment. It aimed to assess both behavioural patterns and perceptions that may influence the child's health and well-being. As shown in Appendix B, the instrument was developed by GicO study experts, who grouped the items into nine domains: (1) physical activity and sport, (2) digital environment at home, (3) eating habits, (4) food access, (5) sustainable eating, (6) nature and environmental contact, (7) sleep quality, (8) mental well-being, and (9) sociodemographic and health status. For illustration, Appendix C (Table 7.1) presents two anonymised parental responses, exemplifying the data structure organised according to these nine domains.

The resulting dataset comprises responses from 38 parents and contains a total of **286 columns**. In the dataset, each column corresponds to a specific questionnaire item that is, a question asked of the respondents, while each row represents the full set of answers provided by an individual parent.

Given the large number of columns, we first summarised variable types at the domain level in Table 2.3. Concretely, we opened the raw file and, domain by domain, inspected the response formats and mapped them to their corresponding variable types, producing a concise overview for easy reference. The table is organised as follows: the first column lists the questionnaire domain; the second column enumerates the typical response formats used by the items within that domain; and the third column reports the variable types those responses generate in the dataset. While the full wording of each question can be consulted in Appendix B, for data analysis, the decisive element is the response itself. Accordingly, we summarise the most common response formats and their resulting variable types to make explicit what we handled in the downstream analyses. In detail, the dataset includes a mix of answer formats: binary options (e.g., Yes/No), ordinal scales (e.g., frequency or agreement Likert scales[2]), discrete numeric values (e.g., minutes, times per week), and open-text fields (e.g., types of foods or physical activities), which reflect the underlying questionnaire design.

Table 2.3: Variable types and response formats by questionnaire domain.

Domain	Response formats	Variable types
1. Physical activity and sport	Yes/No, frequency, duration, free text	Binary, discrete/continuous numeric, ordinal, free text
2. Digital environment at home	Yes/No, age, usage time, apps/websites	Binary, discrete numeric, ordinal, free text
3. Eating habits	Meal counts, food frequency, Yes/No, free text	Discrete numeric, ordinal categorical, binary, free text
4. Food access	Frequency, Likert 1–6, motivation free text, Yes/No	Ordinal categorical, Likert scale, binary, free text
5. Sustainable eating	Agreement scale from 1 to 6	Ordinal categorical (Likert)
6. Nature and environmental contact	Distance (metres), quality rating, visit frequency	Discrete numeric, ordinal categorical
7. Sleep quality	Sleep schedules, Sleep-related Likert items (0–3)	Ordinal categorical
8. Mental well-being	Psychological states Likert items (0–3)	Ordinal categorical
9. Sociodemographics and health status	Sex, age, height/weight, education, diagnoses	Categorical (nominal/ordinal), discrete numeric, binary, free text

2.4.2 Body composition data

The body composition dataset was derived from physical assessments conducted at the Cardiorespiratory and Metabolic Unit (CRMU) of the Faculty of Sport Sciences and Physical Education, University of Coimbra (FCDEF-UC) [4], using standardised and validated procedures. Data were collected through direct measurements using anthropometry and bioelectrical impedance analysis (BIA), with equipment such as the InBody 770 [7], which provides high-precision estimates of lean mass, fat mass, and total body water, among others.

The detailed data structure for the *Body Composition* dataset is documented in Appendix D, where we provide raw excerpts (see Table 7.2). The raw file delivered for

[2]A Likert scale is a type of response scale used to rate levels of agreement or frequency [6].

analysis comprised **234 columns**, spanning primary bioimpedance outputs, device–derived indices, and reference intervals produced by the InBody equipment.

Measurements include continuous body-composition indicators (e.g., body fat mass and percentage, fat-free mass, skeletal muscle/segmental lean mass and fat mass) and water distribution (total, intracellular, and extracellular). Several of these variables are accompanied by an *observation–specific* reference interval reported by the device. Concretely, for a variable v the dataset records a triplet (v, L_v, U_v), where L_v and U_v denote the *Lower* and *Upper* limits of the device's normal range, conditioned on the child's characteristics.

Given the extension of the raw file (234 columns), Table 2.4 presents only a small, representative subset of features; the full set is placed into Appendix D for completeness; the table is three columns list, respectively, the *Variable Group*, *Example Variables*, and their *Type*. The examples highlighted here were chosen because we considered they are the most interpretable to the reader and most informative of underlying bioimpedance constructs (adiposity, lean mass, and fluid compartments). The grouping under "Demographic Information" (age, sex, height, weight) is a logical organisation introduced for readability, these fields appear dispersed in the original file and were not pre-grouped by the GicO study.

Table 2.4: Variable types in the body composition dataset (illustrative subset; full listing in Appendix D).

Variable Group	**Example Variables**	**Type**
1. Demographic Information	Age, Height, Weight, Sex	Continuous, Continuous,Continuous, Categorical
2. Bioimpedance Measures	Body Fat Mass, Fat-Free Mass, Body Mass Index, Segmental Lean Mass, Segmental Fat Mass, Total Body Water, Intracellular Water, Extracellular Water, Obesity Degree, Waist–Hip Ratio, Basal Metabolic Rate	Continuous

2.4.3 Cardiorespiratory data

The study delivered a dataset comprising **four** variables, each capturing a physiological indicator of aerobic capacity and cardiovascular endurance. These measures are central to assessing children's aerobic fitness and overall cardiorespiratory health. Data were collected by the study team using an incremental cycle-ergometer protocol with continuous respiratory gas analysis and heart-rate monitoring.

Table 2.5 lists the variable names, descriptions, and types, and Appendix E (Table 7.3) provides raw excerpts illustrating individual test outputs. Subsequently, for the data analysis in this work we retained only VO_2 max, reflecting guidance from GicO domain experts that prioritised the variable as the most informative marker for cardiorespiratory endurance.

Table 2.5: Cardiorespiratory assessment variables and their types.

Variable	Description	Type
VO_2 max (L)	Absolute oxygen consumption	Continuous
VO_2 max (mL/Kg/min)	Relative oxygen consumption per kilogram of body weight	Continuous (normalised)
Power output (W)	Maximum workload achieved during the test	Continuous
Time to exhaustion (s)	Total duration of the incremental exercise test until voluntary exhaustion	Continuous (time-based)

2.4.4 Functional data

The GicO study delivered a functional fitness dataset for use in this work, capturing children's motor performance through a battery of validated tests for paediatric populations. The dataset comprises **five features**: 20 m sprint time, squat–jump height, right–hand grip strength, left–hand grip strength, and seat–and–reach (cm). Consistent with the body composition and cardiorespiratory datasets, all variables were collected by the study team under standardised protocols. Table 2.6 lists the variable names, descriptions, and types, while Appendix F (Table 7.4) provides a raw-data excerpt illustrating the structure and coding of the first two entries.

Table 2.6: Functional assessment variables and their types.

Variable	Description	Type
20-metre sprint	Sprint time over a 20-metre distance, measured in seconds, assessing speed and acceleration	Time-based (continuous)
Handgrip strength (right and left hand)	Upper body muscular strength measured in kilograms using a handheld dynamometer	Continuous
Squat jump (SJ)	Explosive lower limb strength assessed through a vertical jump without prior countermovement	Continuous
Countermovement jump (CMJ)	Jump height obtained with preparatory knee flexion, reflecting use of the stretch–shortening cycle	Continuous
Sit-and-reach	Trunk and hamstring flexibility measured in centimetres using the sit-and-reach test	Continuous

2.4.5 Drawings data

Children's drawings were used as a child-friendly instrument to uncover perceptions of daily behaviours and healthy choices, complementing the survey and fitness measures with a visual, low-literacy modality. Rather than receiving the original drawings, we obtained a pre-processed dataset that encodes each drawing into structured categories suitable for GicO goals (e.g., presence of healthy foods, activities, social context, and setting). This drawing-based "questionnaire" captures what children choose to depict and how they prioritise elements related to the GicO Study. In total, 34 children produced drawing sets in response to four guiding prompts:

1. *What is your favourite food?*
2. *What do you think makes you grow strong and healthy?*
3. *What do you enjoy doing the most?*
4. *Who would be the best superhero to represent the study?*

GicO experts examined each drawing set and first recorded a short textual description of the depicted content (e.g., "spaghetti bolognese", "gym", "watching cartoons", "super-fruit"). Next, they induced a study-specific set of categories directly from the corpus of drawings, selecting those most representative and meaningful to the study aims (e.g., *Water*, *Soup*, *Vegetables*, *Fruits*, *Cereals & derivatives*, *Meat, fish & eggs*, *Exercise*, etc...). Each drawing was then *one-hot encoded*: for every category, a binary indicator was assigned (1 if the concept appears in the drawing; 0 otherwise). The result is a coded dataset.

An illustrative example is shown in Figure 2.8, depicting a child's response to "What makes you grow strong and healthy?" through gymnastics. A raw excerpt of the encoded dataset (first two entries) is provided in Appendix G (Table 7.5) to clarify the final structure used downstream.

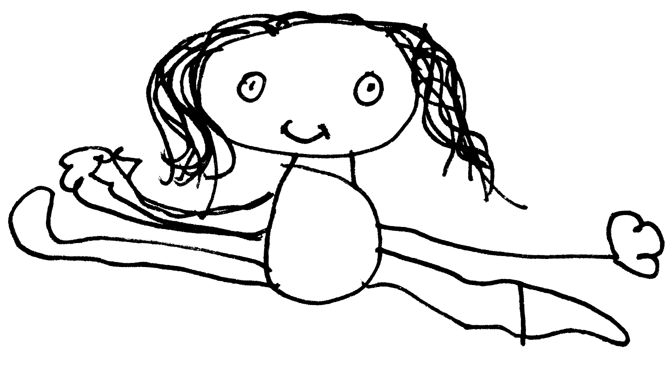


Figure 2.8: Example drawing for the prompt "What makes you grow strong and healthy?", depicting gymnastics as the response.

2.5 Related work

Building on the previous sections, where we detailed five datasets spanning body composition, cardiorespiratory, functional fitness, parental questionnaires, and drawings, we now review healthcare AI methods that underpin our analysis following the objectives of this work. The focus is not on multi-source integration or obesity-specific studies, but on techniques suited to our setting: high dimensionality (feature selection and dimensionality reduction), uncovering latent structure (clustering), handling questionnaire-style variables and mixed types, and maintaining robustness with small samples (n=34 children).

Accordingly, we emphasise prior work on: (i) methods for questionnaire and mixed-type data; (ii) strategies for high-dimensional problems, including regularisation and feature selection; (iii) unsupervised learning for subgroup discovery (clustering); and (iv) construction of composite indices for health assessment. These themes collectively motivate the modelling decisions applied to the datasets introduced above.

2.5.1 Dimensionality Reduction for High-Dimensional Health Data and Feature Selection

The application of AI in healthcare has evolved rapidly, addressing challenges such as the early detection of chronic diseases and the identification of lifestyle-related risk factors [8]. Health-related questionnaires often contain numerous variables covering behaviour, lifestyle, and socio-environmental aspects. To address this high dimensionality, Principal Component Analysis (PCA) is a widely used method that transforms correlated variables into uncorrelated principal components while retaining maximum variance [9, 10, 11]. In previous studies, PCA has been effectively combined with clustering algorithms to improve patient subgroup identification [12, 13].

However, PCA lacks variable selection capability and often results in components that are difficult to interpret, which is a limitation when transparency is crucial for public health recommendations. For non-linear relationships, alternatives like t-SNE (t-Distributed Stochastic Neighbour Embedding) and MDS (Multidimensional Scaling) have shown success in structuring complex datasets [14, 11].

Beyond projection-based approaches such as PCA, regularisation offers a complementary route to dimensionality reduction with stronger interpretability. Lasso (Least Absolute Shrinkage and Selection Operator) encourages exact zero coefficients, yielding compact, human-readable models that foreground a small set of predictors rather than opaque linear combinations [15].

In healthcare AI, where high-dimensional, correlated variables are common across domains like genomics, radiomics, electronic health records and multi-item questionnaires, Lasso and related variants (e.g., elastic net) are widely reported to stabilise estimation under multicollinearity while improving transparency and clinical auditability. Recent work with patient-reported outcomes (PROMs[3]) illustrates this clearly: a *Clinical Cancer Research* study used Lasso to narrow baseline characteristics into a small set of indicators for survival modelling [17]. This line of work positions feature selection via regularisation as a practical alternative to projection methods (PCA) when the goal is not only variance control but also the identification of a minimal, interpretable subset of predictors linked to health outcomes.

2.5.2 Clustering Approaches in Health Risk Profiling

Unsupervised clustering is widely used to derive population segments from health indicators, with K-means and hierarchical methods among the most common choices [18, 13].

For small cohorts, such as the GicO sample, hierarchical clustering is often preferable because it (i) does not require pre-specifying the number of clusters, (ii) yields an interpretable dendrogram that experts can inspect at multiple resolutions, and (iii) is less sensitive to random initialisation than centroid-based algorithms. In practice, linkage choices (e.g., Ward's method) makes hierarchical procedures adaptable to heterogeneous health data; methodological guides recommend it for small-to-medium datasets, and small-scale healthcare pilots have successfully applied it to derive actionable subgroups [19, 20].

When samples are larger and predominantly continuous, dimensionality reduction followed by centroid-based clustering remains effective; indeed, previous studies have combined PCA with K-means to segment patients along latent constructs [10]. Taken together, these practices position hierarchical clustering as a robust, transparent first-line approach for small-sample health profiling, with PCA+K-means serving as a complementary option in higher-n regimes.

2.5.3 Composite Indicators

Given the scope and heterogeneity of questionnaire items applied to parents in GicO Study, spanning domains such as physical activity, diet quality, sleep, and socio- environmental context, it is reasonable to study the instrument's structure and consider compressing item-level detail into domain-level indices. The OECD *Handbook on Constructing Composite Indicators* provides a well-established blueprint for this task, emphasising principled variable selection, treatment of missing data, normalisation of scales, explicit weighting, and transparent aggregation rules [21]. Framed this way, composite

[3]Validated questionnaires capturing patients' self-assessments of symptoms, functioning and health-related quality of life [16].

indicators act as dimensionality-reduction devices that preserve substantive meaning: they translate many correlated items into a few interpretable domain scores, facilitate comparisons across respondents, and offer inputs that are readily audited by experts. Best practice further recommends sensitivity analyses (e.g., alternative normalisations or weights) and construct validation (e.g., convergence with related measures), helping ensure that any derived indices remain robust, communicable, and aligned with the questionnaire's conceptual design.

2.6 Summary

The reviewed literature helps to identify which AI models and techniques are well-suited to datasets like those in the GicO study, small cohort, multi-source, mixed-type, and high-dimensional, and, therefore, to the nature, scope, and objectives of this thesis. Prior work consistently supports dimensionality reduction (e.g., PCA) and unsupervised subgroup discovery (e.g., hierarchical clustering) for health profiling, and it highlights the value of interpretable feature selection via Lasso regularisation. However, relatively few studies address genuinely multi-source, multidimensional settings, and fewer still combine Lasso-based feature selection with hierarchical clustering to yield transparent, audit-ready profiles.

Guided by these insights, this work focuses on (i) constructing domain-level composite indicators to compress parental questionnaire complexity while preserving meaning, (ii) applying Lasso regularisation to identify a minimal set of stable, interpretable predictors, and (iii) using hierarchical clustering, well matched to small n, to reveal behaviourally meaningful subgroups of children. Together, these choices aim to balance interpretability with robustness, aligning with early screening and prevention goals for the specific cohort contexts.

3 Exploratory Data Analysis

This chapter describes the methodological steps followed in exploratory data analysis (EDA) performed on the data collected in the GicO study [2]. The aim was to extract insights related to childhood obesity risk by applying data-driven techniques grounded in exploratory analysis, feature engineering, and unsupervised machine learning models.

This chapter documents the exploratory data analysis (EDA) that prepares the multi-source dataset for the modelling work in Chapter 4. With a small cohort and heterogeneous inputs, our aim is to make the data analysis pipeline clear and reproducible while keeping modelling assumptions minimal.

We begin with *data cleaning*, ensuring each source is internally consistent and linkable. This includes aligning identifiers across files, addressing missing or duplicated entries, and applying simple checks to reduce obvious errors before any modelling decisions are made.

Next, we move to *data preprocessing*, where variables are made comparable and analysis-ready. Steps include standard transformations (e.g., scaling/encoding), handling mixed response formats, and constructing a small number of interpretable derived features to capture meaningful patterns compactly. We also perform *feature selection* to reduce redundancy and focus subsequent models on the most informative signals, and *feature engineering* to summarise complex inputs into clearer, domain-aligned representations.

Finally, we use *data visualisation* to characterise the dataset and guide downstream choices. Univariate views describe distributions and potential anomalies, while simple bivariate explorations highlight relationships and overlap between variables. The chapter closes by summarising what these findings imply for the next stage, motivating dimensionality reduction and structure-discovery methods introduced in Chapter 4.

Software and hardware. All analyses were developed in PyCharm[1] 2023.3.7 (Professional Edition) using a dedicated Python[2] 3.11 virtual environment. Core libraries were: pandas 2.2.3, scikit-learn[3] 1.5.2, SciPy 1.14.1, matplotlib 3.9.2, seaborn 0.13.2, and plotly 5.24.1. Computations ran on a Windows 11 Home laptop equipped with an AMD Ryzen 7 5800H (3.20 GHz) and 16 GB RAM.

[1] `https://www.jetbrains.com/pycharm/`
[2] `https://www.python.org/`
[3] `https://scikit-learn.org/`

3.1 Data cleaning

A preliminary exploration of the datasets revealed missing values whose distribution and nature vary across sources. In data cleaning, we aligned children's identifiers to ensure referential integrity (between parental questionnaire and body composition datasets), removed duplicates, audited and flagged missingness, reconciled dual-parent questionnaire entries (father and mother) into a single child-level record, and applied basic range/type/plausibility checks, including outlier screening, to address data-entry errors before analysis.

Regarding the parental questionnaire dataset shows high completeness for closed-ended questions, which constitute the majority. In contrast, several open-ended or conditional follow-up questions exhibit missing values, often reflecting non-response or irrelevance to the respondent's context. Table 3.1 summarises these missing entries, identifying questions with at least one missing response.

Table 3.1: Questions with Missing Values in the Parental Questionnaire Dataset.

Null Count	Column
19	If answered yes: [How many times per week?]
19	If answered yes: [How many hours per session/training?]
19	If answered yes, which sport(s)?
38	Which digital devices do you have at home? [Other]
32	Which ones are usually used in the bedroom? [Other]
1	Age at first use: Television
1	Age at first use: Computer
14	Age at first use: Video game console
6	Age at first use: Tablet
38	Activities done on digital devices
38	Social media and most used apps/sites
38	Social media and apps
38	Most watched channels and programmes
38	Games played most often
4	Seasonal foods most consumed
37	Prefer not to answer [Comment]
38	Education level completed: [Other]
27	Currently working? [Yes][Comment]
38	Currently working? [Unemployed since:][Comment]
38	Currently working? [Retired. Last profession?][Comment]
35	Diagnosed conditions: [Other]
37	Family medical history (ascendant): [Other]
37	Family medical history (descendant): [Other]
18	Open comments by participant

Missing data were also observed in the body composition dataset, particularly in specific clinical features such as normal range limits or derived cardiovascular metrics.

Table 3.2 lists the variables with missing values and their corresponding counts. Notably, some cardiovascular features (e.g., blood pressure and derived indices) were entirely missing across participants and were excluded from the analysis due to lack of information and irrelevance to the core objectives.

Table 3.2: Variables with Missing Values in the Body Composition Dataset.

Null Count	Feature
1	18. TBW (Total Body Water)
2	19. Lower Limit (TBW Normal Range)
2	20. Upper Limit (TBW Normal Range)
3	29. Upper Limit (Protein Normal Range)
1	31. Lower Limit (Minerals Normal Range)
1	32. Upper Limit (Minerals Normal Range)
2	33. BFM (Body Fat Mass)
1	35. Upper Limit (BFM Normal Range)
1	36. SLM (Soft Lean Mass)
34	220–225. Blood pressure and derived cardiovascular metrics
34	229. Medical History
34	230. Group

The other datasets had very few missing values. The functional dataset contains only one missing entry for the 20m sprint test. Similarly, one record was missing in the drawings dataset for the prompt "What do you enjoy doing the most?". The cardiorespiratory dataset was complete with no missing values.

During the data loading process, placeholder characters such as "X" or "-" were systematically replaced with `NaN`. Variables with missing values in all rows were removed entirely. For numerical features in the body composition and functional datasets, **missing values were imputed using the mean**. In contrast, free-text fields from the questionnaire, often optional and sparsely filled, were excluded from analysis, as the study does not include natural language processing.

A few extreme values were reviewed in the cleaning phase, including one respondent who reported over 50 hours of weekly physical activity due to triathlon training. Although preserved, such records were treated cautiously during statistical analysis.

Additionally, equipment control variables were removed, and decimal character formats were standardised (e.g., replacing commas with periods) to ensure consistent numeric parsing.

3.2 Data preprocessing

This section describes the transformation of raw datasets into structured and meaningful features that enhance the interpretability and predictive performance of statistical and machine learning models. These transformations included expert-guided filtering, index construction, dimensionality reduction, and regularised feature selection methods. The process was adapted to the specific characteristics of each data source, namely, the parental questionnaire, body composition data, cardiorespiratory, functional, and children's drawings.

Expert-guided feature selection Given the relatively small sample size and the clinical relevance of many physiological variables, feature inclusion was guided by domain experts from the Faculty of Sport Sciences and Physical Education and the Center for Neuroscience and Cell Biology (CNC-UC). Based on this input, raw segmental impedance measurements (e.g., for limbs) and electrical bioimpedance features were excluded due to limited reliability or redundant information. Instead, integrative indicators of whole body composition were retained. An exception was made for the variable *'207. 50kHz-Whole Body Phase Angle'*, which has recognised predictive value in body composition dynamics and was therefore preserved [22].

Parental Questionnaire. When both parents responded, a single record was retained per child (selected at random without preference) to avoid duplication. On the other hand, given the complexity and volume of the questionnaire (over 280 variables), treating each item independently would not result in interpretable or meaningful features. To facilitate analysis, all questionnaire items were grouped into nine domains mentioned in Section 2.4.1, these are: (1) physical activity and sport, (2) digital environment at home, (3) eating habits, (4) food access, (5) sustainable eating, (6) nature and environmental contact, (7) sleep quality, (8) mental well-being, and (9) sociodemographic and health status. This organisation follows the original structure of the questionnaire, which was designed around these nine domains, and therefore, no additional regrouping or new constructs were introduced. Maintaining this predefined framework ensures conceptual consistency with the instrument's design, enables targeted exploratory analysis, and supports the identification of patterns or risk factors relevant to childhood obesity. Instead, variables were grouped thematically and consolidated into composite indices (see Section 3.2.2). This approach enabled us to reduce dimensionality while preserving the latent constructs originally targeted by the questionnaire design. Although the process of calculating these indices was challenging, it simplified downstream analysis and enhanced interpretability.

Body Composition. For body composition data, reference ranges were used to classify variables into categories (e.g., LOW, NORMAL, HIGH). These categories were used for exploratory profiling but excluded from modelling to avoid redundancy. Expert consultation also led to the removal of anthropometric estimates such as arm circumference. A subset of whole body features, including fat mass, lean mass, water content, and metabolic indicators, was retained. Additionally, new features such as BMI percentile and BMI category were derived using child-specific reference curves provided by the Portuguese Directorate-General of Health [23].

Functional and Cardiorespiratory Data. No derived variables were constructed for these datasets. The raw measurements were retained and directly used in multivariate analysis.

Children's Drawings. Free-text descriptions and thematic codes (eg., Vegetables, Ultra-processed food, Nature/Environment) extracted from children's drawings were also transformed into indices. This process mirrored the questionnaire logic: healthy elements contributed positively to the score (e.g., fruits, exercise), while unhealthy elements reduced it (e.g., fast food, sedentary activities). The result was a set of structured variables that captured the child's implicit perception of health and nutrition.

3.2.1 Feature transformation

The body composition dataset contains continuous variables expressed in heterogeneous units (e.g., kilograms, centimetres, calories). To ensure comparability across features and to stabilise optimisation in downstream models, we standardised all continuous features via *z-score scaling* (scikit-learn's `StandardScaler`). This step is particularly important for distance-based methods (e.g., clustering), regularised linear models (e.g., Lasso), where disparate feature scales can distort distances, penalty terms, or learning dynamics.

Method (StandardScaler). For each feature x, `StandardScaler` computes the sample mean μ and sample standard deviation σ across the whole dataset used in this step, and transforms each value to

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

This recentres the feature to mean ≈ 0 and rescales it to unit variance (≈ 1). The transformation preserves the order of observations and linear relationships; it strictly shifts and scales, it does not change the underlying distributional shape.

We applied `StandardScaler` to the body composition dataset after prior cleaning. The transformation was performed *once* on the full dataset, resulting in a matrix where each feature has a mean ≈ 0 and variance ≈ 1. This ensured a balanced contribution of mass, ratio measures, and others to subsequent exploratory analyses and modelling performed on the same data.

3.2.2 Feature engineering

Indices construction from the parental questionnaire

The parental questionnaire was *a priori* organised by the GicO experts into domains (e.g., physical activity, digital environment, dietary habits, food access, sustainable-diet literacy/attitudes, neighbourhood environment, sleep quality, emotional well-being, and sociodemographics), with items grouped by purpose and measurement intent. We explicitly leverage this domain structure to handle high dimensionality while preserving each domain's evaluative meaning: rather than modelling hundreds of single items, we construct one composite index per domain and analyse at the level of these domain scores.

To condense the questionnaire into an analytically tractable set of features, we created nine composite indices (see Table 3.3), one per behavioural or contextual domain introduced in Section 2.4. Item selection follows the questionnaire sections mapped to validated international tools and GicO's domain project.

Each index aggregates thematically coherent items using predefined scoring rules and, where applicable, normalisation to a common 0–100 scale (higher scores denote healthier or more favourable behaviours).

Calculation methodology. For domain's indices based on standardised instruments, such as domain (7) sleep quality and (8) emotional well-being, we applied the original scoring procedures without modification (Pittsburgh Sleep Quality Index for sleep; DASS-21 for stress, anxiety, and depression). These instruments are widely used and validated, and their items appear as dedicated sections in the parental instrument [24].

For project-specific indices (e.g., PAI, STI, GAI), categorical responses were first translated into numerical quantities (e.g., minutes of activity, MET estimates, weekly frequencies), combined through linear aggregation, and rescaled to 0–100 so that higher values consistently indicate healthier or more protective behaviours.

Overview of constructed indices:

- **PAI** – Physical Activity Index: MET-h/week[4] estimated from frequency and duration of sports and daily activities, following the Compendium of Physical Activities [25].
- **STI** – Screen-Time/Sedentary Time Index: based on daily hours spent on screens or sitting. For adults, risk levels correspond to observational evidence such as that from Stamatakis et al. [26], which found that sedentary time > 8h/day was associated with a 52% increase in all-cause mortality compared to < 4h/day.

[4]MET-h/week stands for Metabolic Equivalent Task hours per week — a standardised unit that combines the intensity of an activity (MET) with its duration. For example, performing a 4 MET activity for 1 hour equals 4 MET-h.

- **DQI** – Diet Quality Index: was constructed by adapting principles from two widely used dietary indices: the Healthy Eating Index–2010 (HEI) [27], which inspired the use of a 0–100 scoring scale based on portion frequency, and the Mediterranean Diet Score (MDS) [28], which provided the conceptual model for separating healthy and unhealthy food groups with differential scoring logic.
- **FIS** – Food Insecurity Score: counts affirmative responses to 14 household food insecurity events [29].
- **SFL** – Sustainable Food Literacy: score scaled to 0–100, constructed from 21 Likert items on food knowledge, skills, and sustainability attitudes, based on the framework by Teng and Chih [30].
- **GAI** – Green Access Index: combines distance to parks, perceived environmental quality, and usage frequency, informed by WHO (World Health Organization) urban green metrics [31].
- **SQI** – Sleep Quality Index: PSQI–PT global score, with higher values indicating poorer sleep quality [32].
- **SAD** – Stress, Anxiety, and Depression subscales derived from DASS-21 [33].
- **SES** – Socioeconomic Status Index: derived from parents' educational attainment, employment status, and household composition.

Table 3.3: Overview of indices constructed from the questionnaire.

Index	Description and components	Scale	Source
PAI	Physical activity (sports + daily activity) in MET-h/week	MET·h/week	[25]
STI	Screen time / Sedentary time (TV, mobile, console, PC) weekday/weekend average	hours/week	[26]
DQI	Weekly food frequencies mapped to HEI-style and MDS scoring	0–100	[27] [28]
FIS	Count of food insecurity events (14 binary items)	0–14	[29]
SFL	21 Likert items on literacy, skills, intention	0–100	[30]
GAI	Distance to green space, quality, frequency of use	0–100	[31]
SQI	Sleep quality (PSQI–PT global score)	0–21	[32]
SAD	Subscales of Stress, Anxiety, Depression (DASS-21)	0–42	[33]
SES	Educational level, employment, and household dependency ratio	0–100	[34]

Distribution and scaling: Most indices were linearly transformed to range from 0 to 100, where higher values reflect healthier, more sustainable or more protective profiles. Exceptions include the FIS (count from 0 to 14), the SQI (score from 0 to 21), and the DASS–21 subscales (0–42 each).

Example of Index Construction: Physical Activity Index (PAI): To illustrate the process of transforming questionnaire responses into meaningful quantitative indicators, we present here the construction of the Physical Activity Index (PAI). This index captures the weekly physical activity level of parents based on two distinct sources of information extracted from the questionnaire:

- **Organised Sports Practice (Section 1.1):** Respondents who answered *"Yes"* to practicing sports provided two quantitative inputs:
  - Frequency (*times per week*)
  - Duration per session (*hours*)

  Weekly activity minutes were calculated as:

$$\text{Weekly Minutes} = \text{frequency} \times \text{duration} \times 60 \tag{3.2}$$

- **Daily Physical Activities (Section 1.2):** Activities such as walking, cycling, gardening, dancing, etc., were reported in ordinal categories (e.g., *"30 to 60 min/ week"*). Each category was mapped to the midpoint value of its interval (e.g., "30–60 min" $\rightarrow$ 45 min).

Conversion to MET-hours/week: To allow for energy-based comparison, the weekly physical activity time was converted to **MET-hours/week**, where MET (Metabolic Equivalent of Task) quantifies the energy cost of an activity relative to resting.

- 1 MET = energy expenditure at rest
- Example values:
  - Walking: 3.3 METs
  - Bicycling: 4.0 METs
  - Running: 8.0–12.0 METs

Formula:

$$\text{MET h/week} = \left(\frac{\text{Minutes per week}}{60}\right) \times \text{MET value} \tag{3.3}$$

Example: A respondent reports:

- Walking: 90 minutes/week $\Rightarrow \frac{90}{60} \times 3.3 = 4.95$ MET h/week
- Cycling: 60 minutes/week $\Rightarrow \frac{60}{60} \times 4.0 = 4.0$ MET h/week

Total PAI = 9.95 MET h/week

This provides a harmonised measure of energy expenditure that facilitates comparisons across individuals with different activity profiles.

Note: Full details on the methodology, scoring rules, and conversion formulas for the other indices are provided in **Appendix H**.

BMI percentile derivation

BMI was complemented with age and sex-specific BMI-for-age percentiles. The body composition dataset provides *age*, *sex*, and *BMI*, which were combined with the official reference curves provided by the Portuguese Directorate-General of Health [23] for ages 2–20 years. The corresponding charts (boys/girls) are included in **Appendix I**.

For each child, the sex-specific chart was selected and the percentile corresponding to the observed BMI at the child's exact age was obtained by interpolation across the age axis. The resulting value was stored as *BMI percentile*. Percentiles were then mapped to standard paediatric weight-status bands: *underweight* (<P5), *healthy* (P5–<P85), *overweight* (P85–<P95) and *obese* (>=P95); values exactly at P85 or P95 were assigned to *overweight* and *obese*, respectively.

This derived variable is used as the supervised target in the feature selection analysis (Section 3.2.2) and as an interpretable descriptor in cluster profiling.

Derived categories from body composition data

Building on the triplet organisation described in the *Body composition dataset*, discrete categories were constructed to express each child's position relative to their observation-specific reference interval. This provides interpretable abstractions of continuous measures while preserving the individualisation inherent to the device-provided bounds.

For any variable v with corresponding lower and upper limits L_v and U_v, the mapping was defined as

$$\mathrm{cat}(v) = \begin{cases} \text{LOW}, & v < L_v, \\ \text{HIGH}, & v > U_v, \\ \text{NORMAL}, & \text{otherwise}, \end{cases} \tag{3.4}$$

This rule was applied systematically to all measures accompanied by device-provided limits. The resulting LOW/NORMAL/HIGH indicators capture personalised deviation from the equipment's expected range and are subsequently used to produce visual summaries and to provide interpretable descriptors for cluster profiling in downstream analysis.

Feature selection from body composition data

To prepare the body composition dataset for subsequent modelling, a supervised feature selection step was undertaken. Body Mass Index (BMI) percentile was chosen as the target variable given its international relevance and its role in classifying children into standardised weight categories, thus serving as a practical proxy for obesity risk.

Lasso regression with an L_1 penalty was used to identify the most predictive variables for BMI percentile. We fixed the regularisation strength at $\alpha = 0.00032$, chosen to yield a compact set of nine non-zero coefficients. This setting reduces dimensionality to the few most relevant features aligned with BMI percentile while shrinking less informative coefficients to zero. Under this α, the model retained nine predictors with non-zero coefficients. These selected features were:

- 21. ICW (Intracellular Water)
- 27. Protein
- 36. SLM (Soft Lean Mass)
- 42. SMM (Skeletal Muscle Mass)
- 137. BMR (Basal Metabolic Rate)
- 146. BCM (Body Cell Mass)
- 154. TBW/FFM (Total Body Water to Fat-Free Mass Ratio)
- 155. FFMI (Fat-Free Mass Index)
- 207. 50kHz-Whole Body Phase Angle

These features were subsequently used in the clustering analysis and serve as core indicators in the profiling of obesity risk among the participating children.

Indices construction from children's drawings

We worked from the *coded* drawings dataset delivered by the GicO experts, who had already pre-processed and encoded each drawing into interpretable categories. Building on their refinements and classifications, we consolidated the codes into broader, composite groupings aligned with three areas: food, health, and physical activity. For each child, we then quantified the presence of elements within these areas (e.g., counts per theme) to obtain structured variables suitable for downstream analysis.

The following total counts were computed:

- **Total Food Elements** (`Food_total`)**:** The sum of all items classified as food.
- **Total Health Elements** (`Health_total`)**:** Elements referring to health-promoting items or behaviours.
- **Total Activity Elements** (`Activity_total`)**:** All representations of physical activity.

In addition, specific subcategories were isolated to identify potentially harmful elements:

- **Ultra-Processed Food Elements** (`Food_ultraprocessed`)**:** Derived from the category `Food_7`, representing items like fast food or sweets.
- **Sedentary Activity Elements** (`Activity_sedentary`)**:** Corresponding to the category `Activity_3`, indicating inactivity or screen-based behaviours.

Based on these raw features, two composite indices were developed to summarise children's implicit health perceptions:

- **Nutrition Perception Index:**

$$\text{Nutrition\_Index} = \text{Health\_total} - \text{Food\_ultraprocessed}$$

  This index represents the balance between positive health-related elements and unhealthy food choices in the drawings.

- **Activity Perception Index:**

$$\text{Activity\_Index} = \text{Activity\_total} - \text{Activity\_sedentary}$$

  This index reflects the net representation of active versus sedentary lifestyles.

3.3 Data visualization

This section provides a concise visual summary of the study data across four domains: parental questionnaire, body composition, cardiorespiratory fitness, and functional performance. Univariate plots describe central tendency, dispersion, and skewness, using continuous histograms alongside categorical views for clarity. Selected body composition variables (e.g., BMI percentile, FFMI, Phase Angle, ICW, SLM, BCM, SMM, Protein, BMR) are shown with paired histograms and boxplots to reveal spread and outliers, followed by baseline distributions for VO_2 max and key functional tests. Correlation heatmaps for questionnaire indices and body composition features then summarise cross-variable structure and potential redundancy, guiding subsequent feature engineering and modelling.

3.3.1 Univariate characterisation of the parental questionnaire dataset

The parental questionnaire provided a diverse set of indicators reflecting lifestyle behaviours, food environment, and psychosocial well-being. To capture this multidimensionality, composite indices were calculated and explored through both continuous and categorical representations.

Physical Activity Index (PAI): The distribution shows strong asymmetry, with the majority of parents reporting low levels of physical activity and only one extreme case exceeding 300 MET h/week (Figure 3.1).

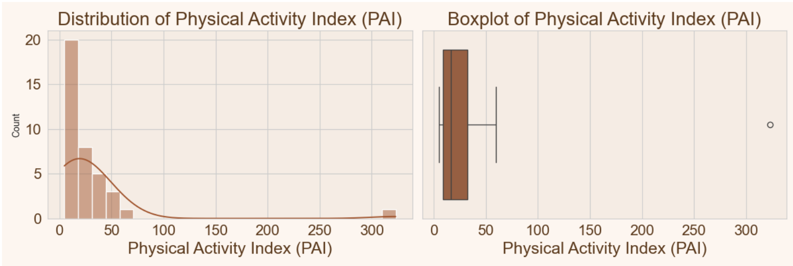


Figure 3.1: Distribution of Physical Activity Index (PAI).

Screen Time–Sedentary Index (STI): Values are concentrated around 40–50 hours/week. The categorical counterpart reinforces this pattern, with the majority of parents classified into high (n 21) and moderate ($n = 13$) sedentary categories (Figure 3.2).

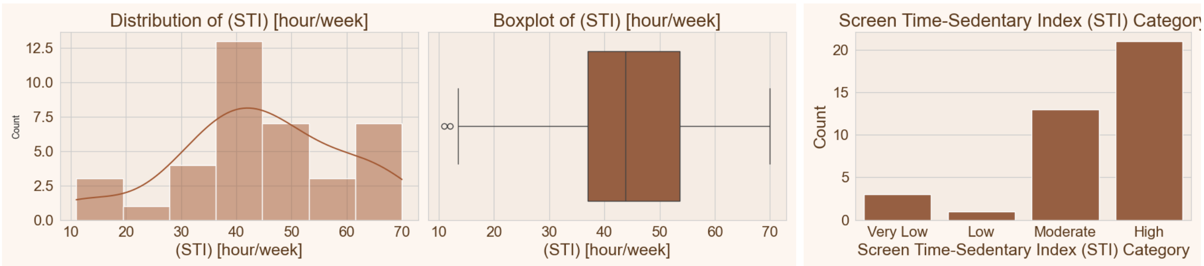


Figure 3.2: Distribution of Screen Time–Sedentary Index (STI).

Diet Quality Index (DQI): The continuous DQI is tightly centred around the mean (50), with a slight tail on the lower end. Categorical analysis revealed a dominant low-quality diet profile ($n = 36$), with almost no participants reaching higher-quality categories (Figure 3.3).

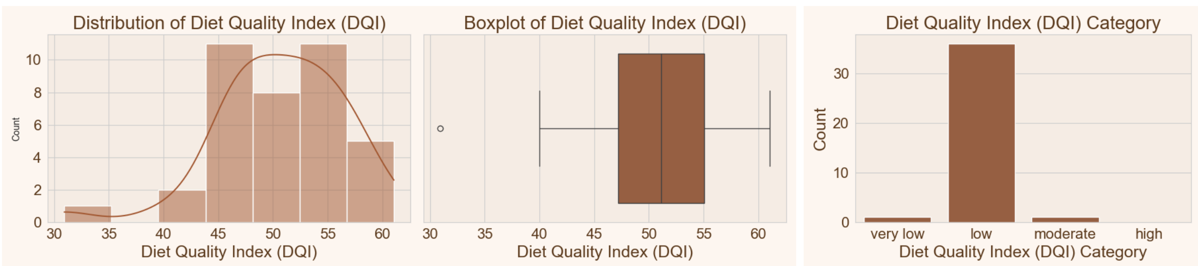


Figure 3.3: Distribution of Diet Quality Index (DQI).

Food Insecurity Score (FIS): The distribution of FIS is highly concentrated near zero, indicating minimal reported insecurity. Categorical results corroborate this, with nearly all families classified as food secure and only one non-severe case reported (Figure 3.4).

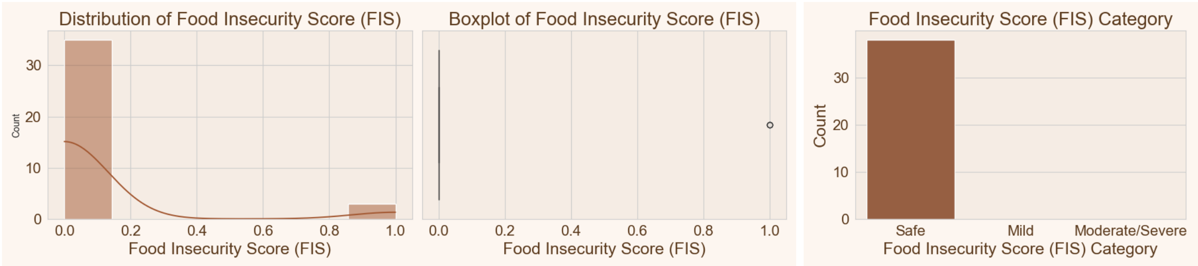


Figure 3.4: Distribution of Food Insecurity Score (FIS).

Sustainable Food Literacy (SFL): The continuous distribution of SFL is skewed toward higher scores, suggesting good general awareness and competencies related to food sustainability. Categorical analysis confirms this, with most participants rated as high ($n = 23$) or good ($n = 12$) (Figure 3.5).

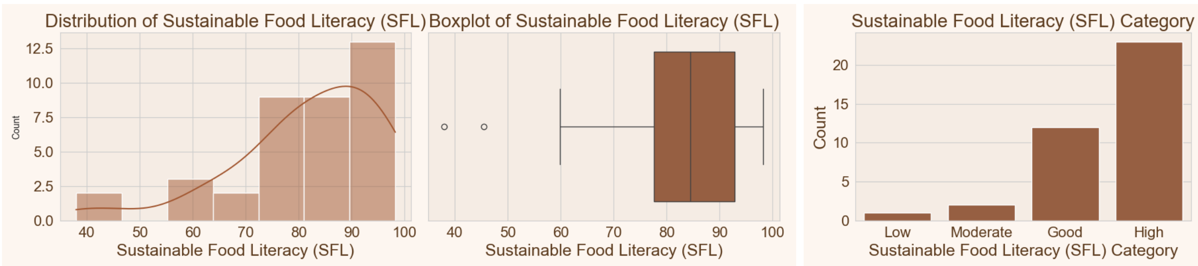


Figure 3.5: Distribution of Sustainable Food Literacy (SFL).

Green Access Index (GAI): Continuous values are spread evenly across the range, with low-amplitude peaks in the mid-range. In the categorical analysis, participants were relatively balanced across low ($n = 16$), moderate ($n = 13$), and good ($n = 9$) levels of perceived access to green areas (Figure 3.6).

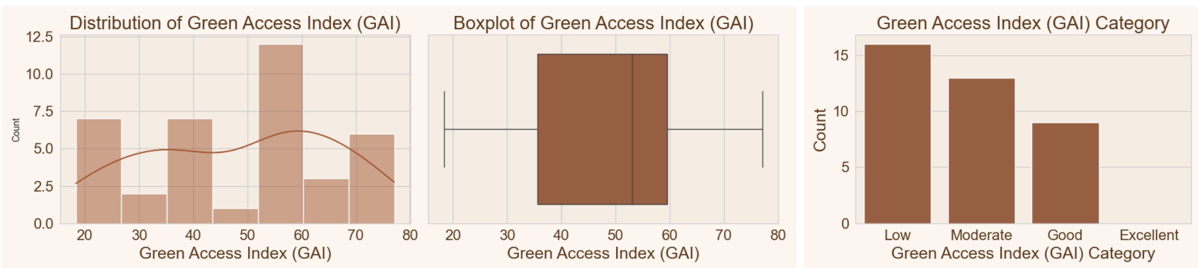


Figure 3.6: Distribution of Green Access Index (GAI).

Sleep Quality Index (SQI): The SQI centres around 6–7 points, with several high-end outliers. Categorically, poor sleep quality was dominant ($n = 30$), suggesting a notable burden of sleep-related difficulties among the parents (Figure 3.7).

Stress, Anxiety, and Depression (SAD): These psychosocial dimensions showed diverse patterns (Figure 3.8):

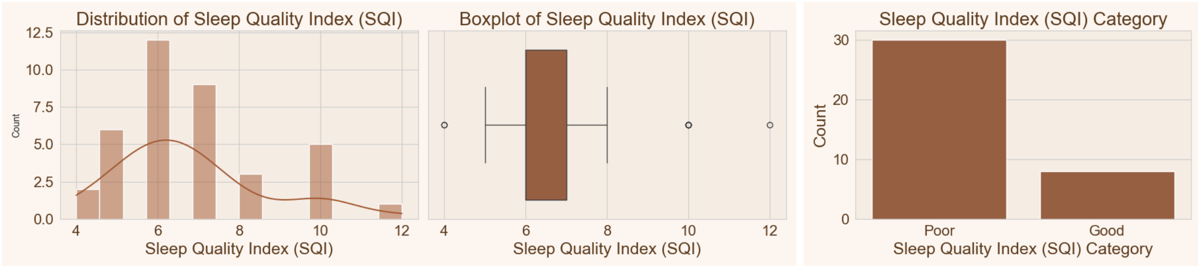


Figure 3.7: Distribution of Sleep Quality Index (SQI).

- **Stress:** Exhibits a wide, fairly even distribution. Categorically, responses range from normal to moderate and severe levels, with 21 parents classified as normal.
- **Anxiety:** Most scores are near zero, with only a few moderate-to-high values. Correspondingly, 35 parents were classified as normal.
- **Depression:** The continuous values are concentrated in the lower range (0–3), with only a few outliers. Categorical results confirm that 31 parents were in the normal range, with 7 mildly affected.

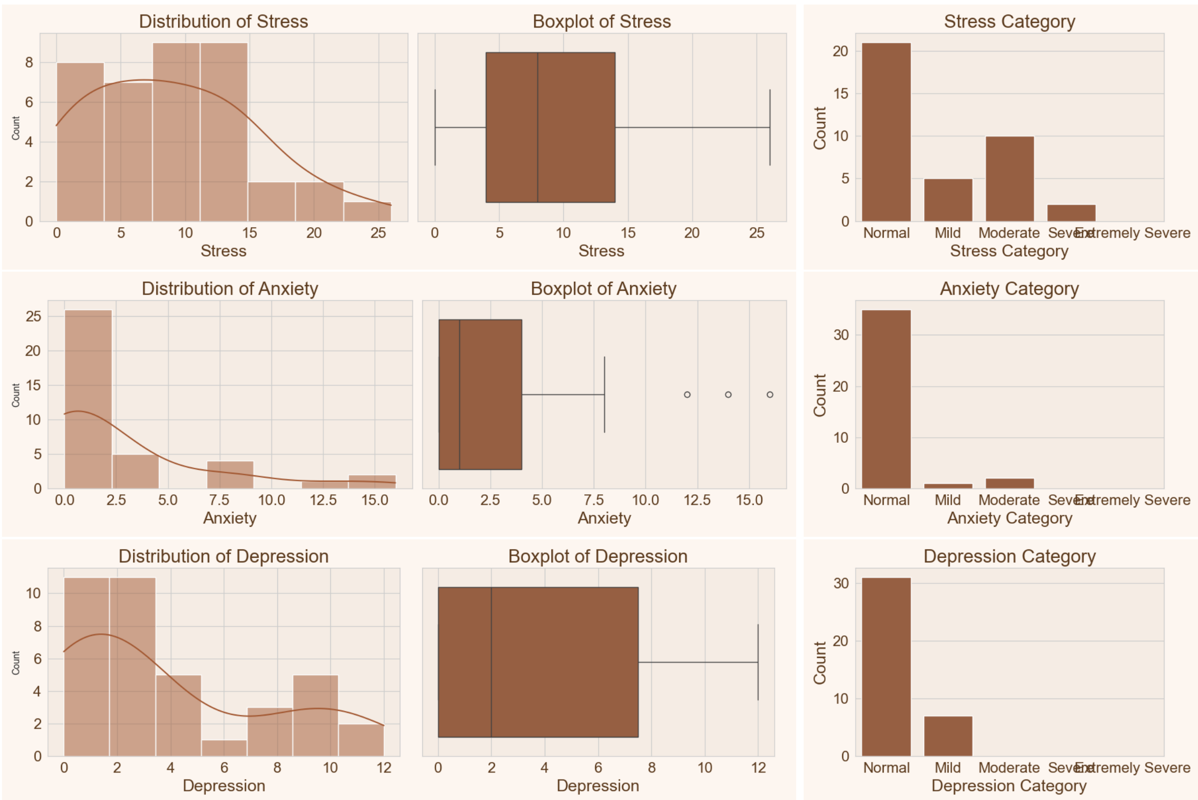


Figure 3.8: Distribution of Stress, Anxiety, and Depression (SAD).

Socio-Economic Status (SES): The SES index shows a distribution centred around 60–70 points, with one extreme low outlier. Categorically, most families fell into upper-middle ($n = 20$) or high ($n = 10$) SES groups, reinforcing the socioeconomic stability of the sample (Figure 3.9).

Figure 3.9: Distribution of Socio-Economic Status (SES).

3.3.2 Univariate characterisation of the body composition dataset

Together, the distributions of the parental indices provide a comprehensive overview of the sample's behavioural, environmental, and emotional landscape in which children are involved. To explore the internal structure and variability of body composition features, univariate analysis was conducted on the subset of variables selected through Lasso regression. These features are particularly relevant for obesity profiling and include classical metrics such as the Body Mass Index (BMI) percentile, as well as advanced indicators from multifrequency bioimpedance, such as the Fat-Free Mass Index (FFMI), Body Cell Mass (BCM), Intracellular Water (ICW), and Phase Angle.

Figures 3.10–3.12 present the distribution and boxplot for each selected variable. Most variables follow a distribution close to normal, with a limited presence of outliers. For instance:

- **BMI Percentile**, although normalised, shows a slight right skew, with a higher density of participants near the upper percentiles, suggesting the presence of overweight tendencies in the population.
- **FFMI (Fat-Free Mass Index)** displays a near-normal distribution, concentrated around the mean with limited outliers, indicating consistent lean mass values across the sample.
- **Phase Angle (50kHz)**, a marker of cellular health[35], presented a distribution profile very similar to that of the Fat-Free Mass Index (FFMI), both displaying approximately normal distributions.
- **ICW (Intracellular Water), SLM (Soft Lean Mass), BCM (Body Cell Mass), and SMM (Skeletal Muscle Mass)** display a broader spread with a small tilt in their distribution, indicating variation in hydration and muscular compartments.
- **Protein and BMR (Basal Metabolic Rate)** present distributions with more uniform spread, pointing to high biological diversity in metabolic processes.

Boxplots complement the histogram analysis by revealing the presence of moderate outliers, especially on the lower end of the FFMI and Phase Angle distributions. These could represent individuals with lower-than-average muscle mass or compromised cellular health.

3.3.3 Univariate characterisation of the cardiorespiratory dataset

Regarding cardiorespiratory performance, VO_2 max at T1 shows a fairly symmetric distribution, with values ranging from approximately 20 to 50 ml/kg/min (Figure 3.13). Most children fall between 30 and 40 ml/kg/min, indicating a moderate level of aerobic fitness for their age group. The boxplot confirms this central pattern, with no extreme outliers observed, suggesting a relatively consistent fitness level across the sample.

3.3.4 Univariate characterisation of the functional dataset

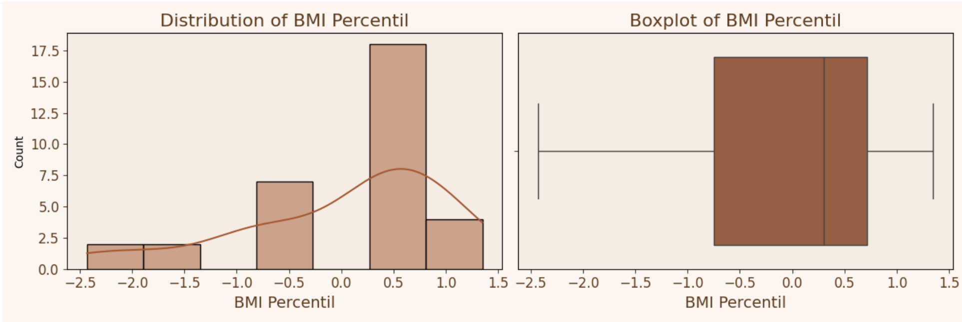
Finally, the distribution of the functional dataset presents a consistent and balanced profile of the children's physical performance at baseline (T1) (Figures 3.14- 3.15). Overall, the group of participants demonstrates normal development patterns across key motor domains:

- **Speed (20M Sprint):** Performance times are mostly concentrated between 4.4 and 4.8 seconds, indicating similar acceleration capacity across children, with no strong outliers.
- **Lower-limb Power (Squat Jump and Countermovement Jump):** shows symmetrical, bell-shaped distributions, with most values falling between 13–18 cm, reflecting homogeneous levels of explosive strength.
- **Upper-limb Strength (Handgrip):** Right and left handgrip strength show comparable distributions centred around 10–12 kg, suggesting balanced upper-body development without marked lateral dominance.
- **Flexibility (Sit-and-Reach):** Scores are distributed from 28 to 40 cm, with a few lower values. Most participants display moderate to good flexibility, with minimal deviation.

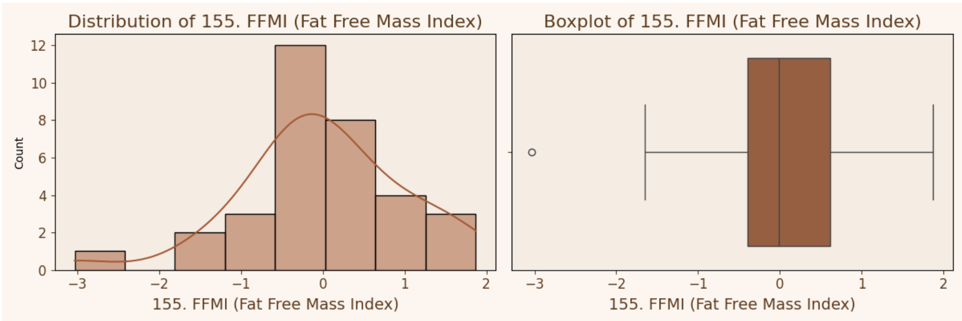
Drawings-coded dataset. We did not run univariate plots for the drawings-coded data. The dataset itself is quite subjective, often sparse or low-cardinality, so it adds little value on its own at this stage. We therefore use it in Chapter 4.

3.3.5 Bivariate distributions

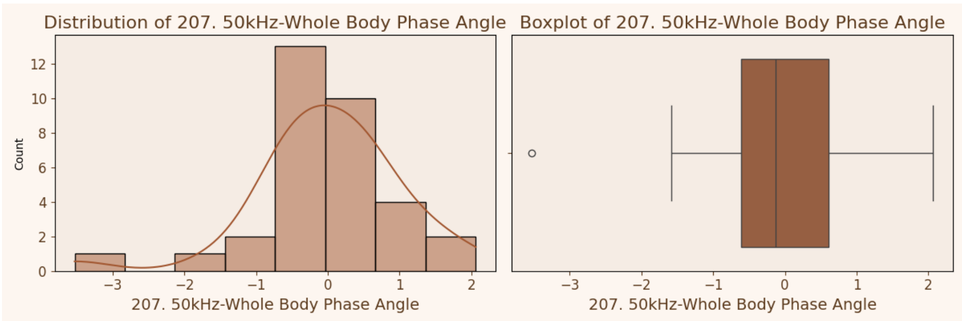
To understand how variables relate to one another across domains, correlation matrices were computed for both questionnaire-derived indices and selected body composition features (Figures 3.16 and 3.17). These heatmaps provide a visual summary of the relationships among variables used in downstream modelling.



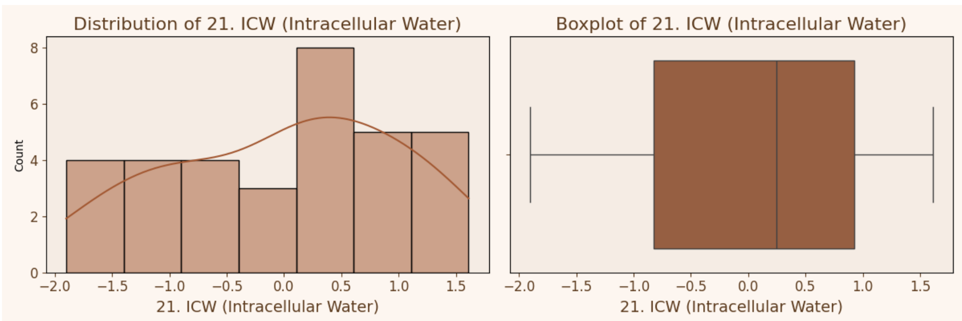
(a) BMI percentile.



(b) FFMI.

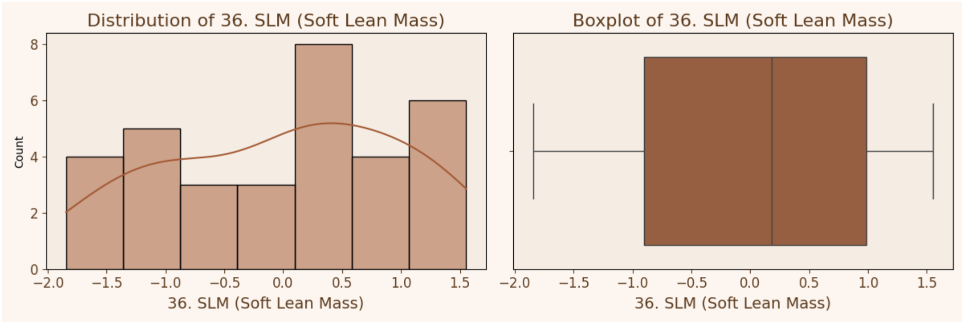


(c) Whole body phase angle (50 kHz).

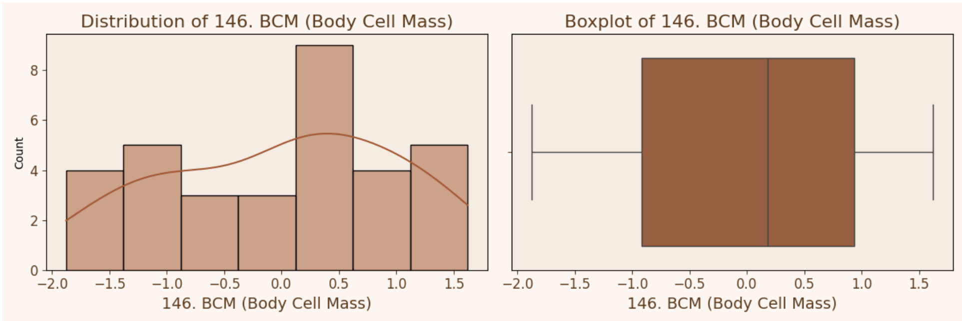


(d) Intracellular water (ICW).

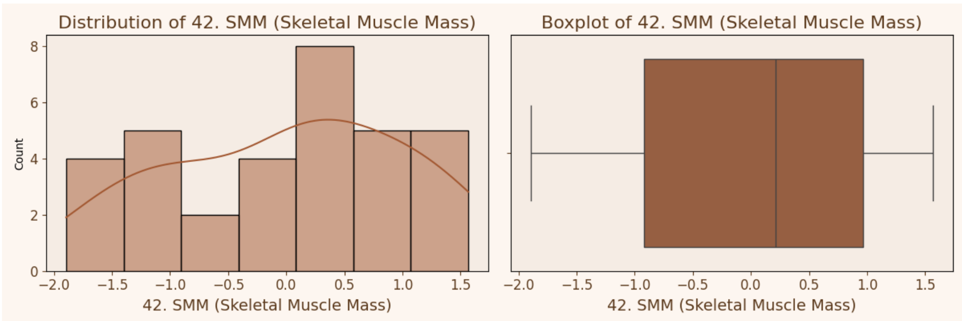
Figure 3.10: Histogram and boxplot of selected body composition variables (Part 1/3).



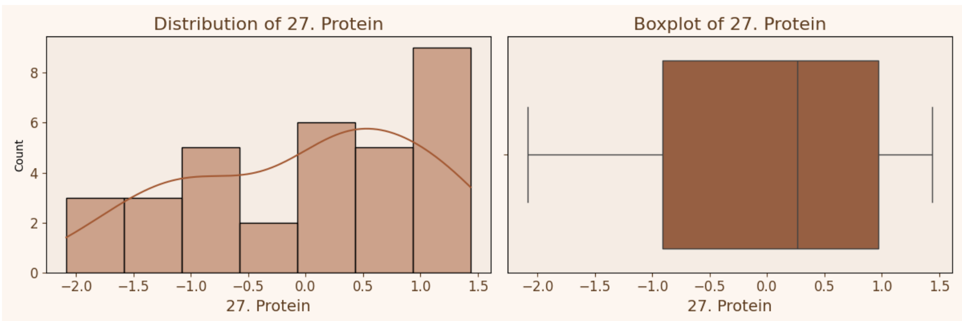
(a) Soft lean mass (SLM).



(b) Body cell mass (BCM).

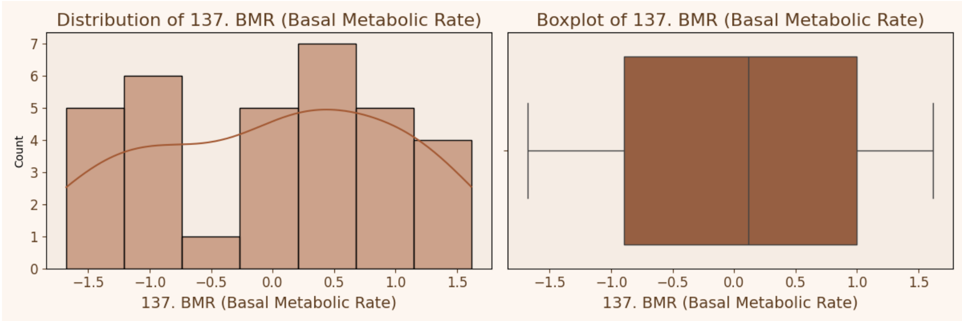


(c) Skeletal muscle mass (SMM).



(d) Protein.

Figure 3.11: Histogram and boxplot of selected body composition variables (Part 2/3).



(a) Basal metabolic rate (BMR).

Figure 3.12: Histogram and boxplot of selected body composition variables (Part 3/3).

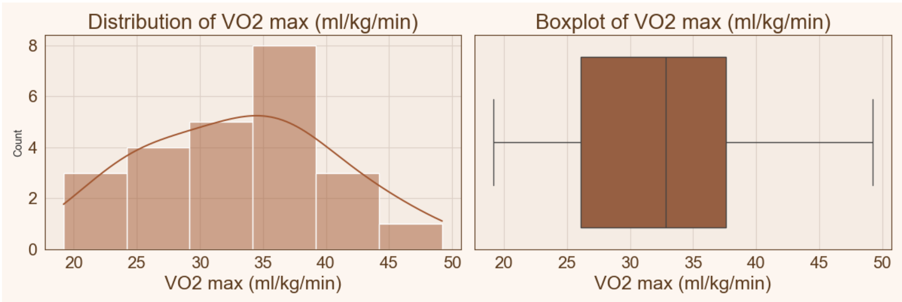
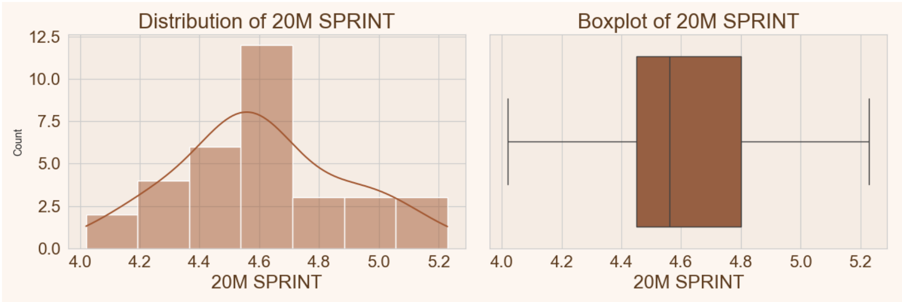


Figure 3.13: Distribution of Cardiorespiratory Fitness.

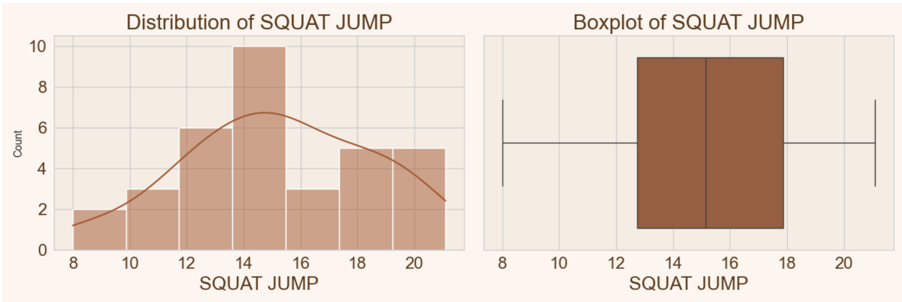
In the **questionnaire-derived indices**, most correlations are low to moderate, indicating relatively distinct constructs. Some expected patterns are observed, for instance, *Stress*, *Anxiety*, and *Depression* are moderately to strongly correlated, reflecting their shared domain of psychosocial health. Likewise, a moderate positive correlation is found between *Diet Quality Index* (*DQI*) and *Sustainable Food Literacy* (*SFL*), suggesting that better nutritional literacy may be associated with healthier dietary habits. No substantial signs of multicollinearity are evident in this set, supporting the inclusion of all indices in further analysis.

In contrast, the **body composition features** display several strong correlations, particularly among variables related to lean tissue (e.g., *SMM*, *SLM*, *BCM*, *ICW*, and *Protein*), with pairwise coefficients often exceeding 0.9. These reflect physiological interdependencies, as these features all describe related aspects of muscular and cellular mass. Similarly, *BMR*, *FFMI*, and *Phase Angle* show consistent alignment with fat-free mass indicators. While these high correlations indicate redundancy and potential multicollinearity, these features were retained due to their individual interpretability and relevance for the GicO project. Their presence enriches the interpretability of derived clusters in later stages.

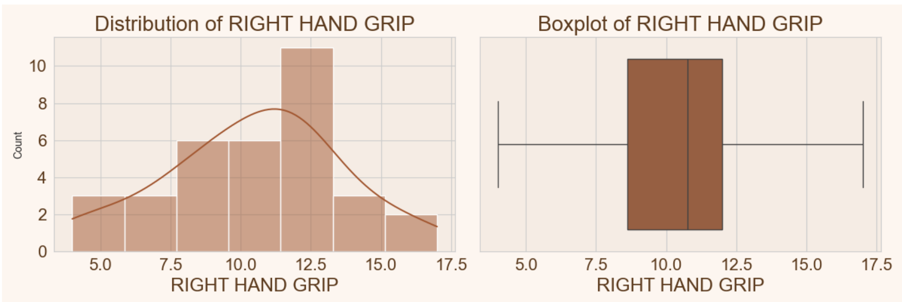
Overall, the correlation analysis helps shift the focus from isolated variables to interconnected patterns, forming a conceptual bridge to the modelling phase that follows.



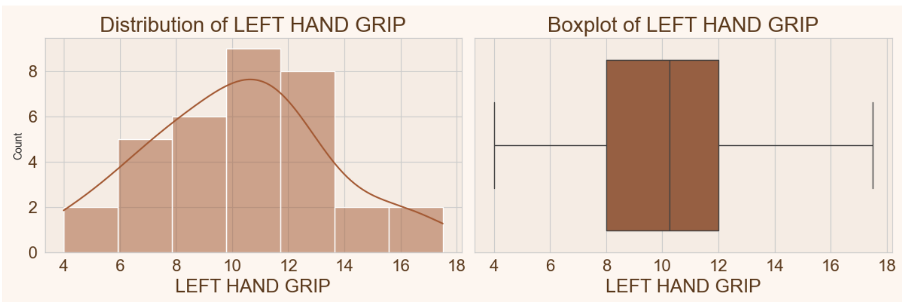
(a) 20 m sprint.



(b) Squat jump.



(c) Right-hand grip.



(d) Left hand grip.

Figure 3.14: Distribution of functional performance (Part 1/2).

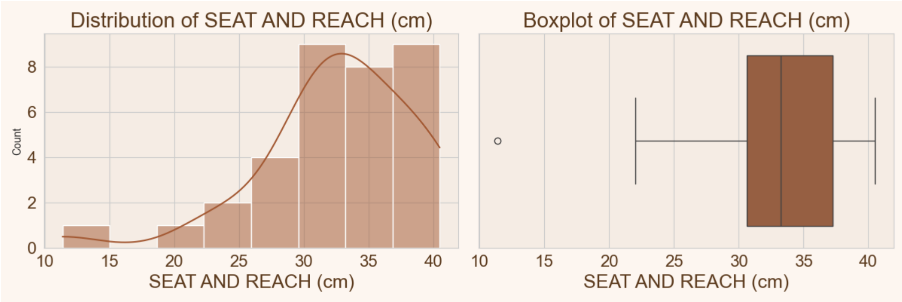


Figure 3.15: Distribution of functional performance (Part 2/2): Seat and reach (cm).

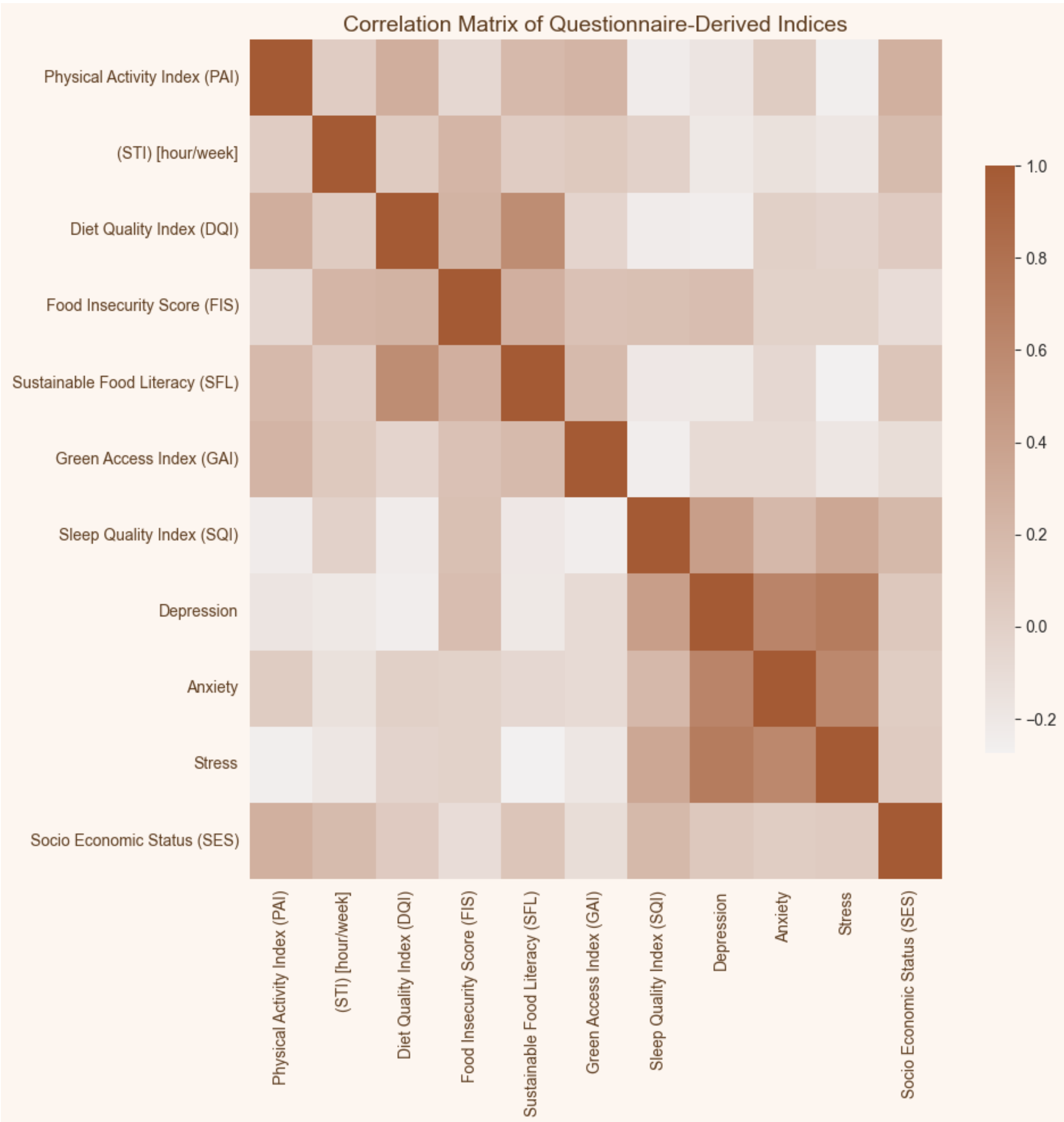


Figure 3.16: Correlation Matrix of Questionnaire-Derived Indices.

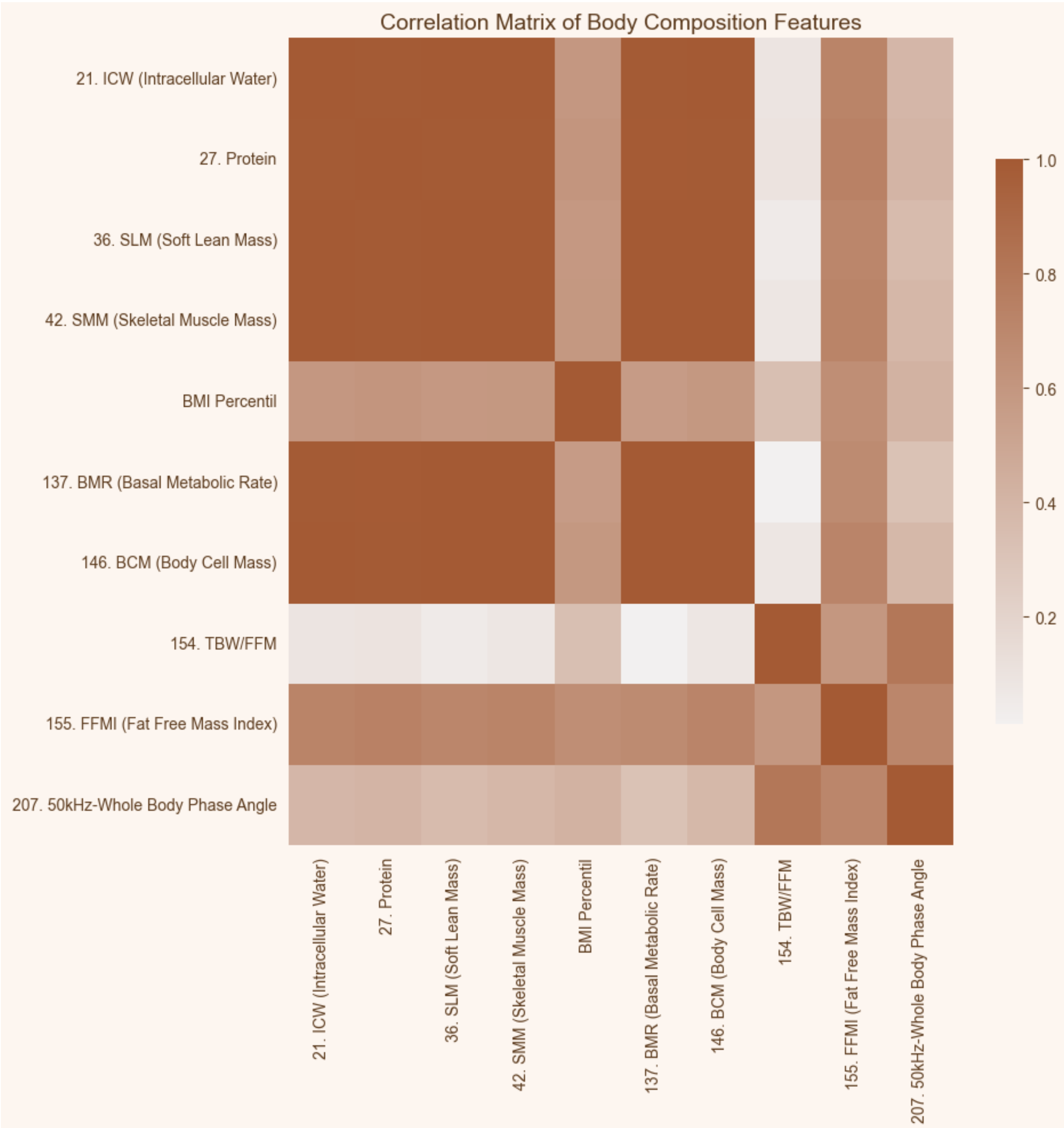


Figure 3.17: Correlation Matrix of Body Composition Features.

3.4 Summary

The exploratory analysis revealed substantial variability across both behavioural and physiological domains. Indices such as the *Fat-Free Mass Index* (*FFMI*), *Screen Time/ Sedentary Index* (*STI*), and *Diet Quality Index* (*DQI*) showed particularly wide distributions, with some participants exhibiting markedly higher or lower values compared to the sample average.

Functional and cardiorespiratory fitness indicators in children (e.g., VO_2 max, sprint time, grip strength) also presented a diverse range of outcomes, suggesting varying levels of physical aptitude. Likewise, psychological measures including *Stress*, *Anxiety*, and *Depression* displayed asymmetric distributions, with a subset of parents reporting more severe symptoms.

This observed heterogeneity spans questionnaire-derived indices, body composition features, and physical assessments, suggesting the presence of underlying group structures or profiles. Such patterns underscore the need for approaches capable of uncovering these latent subgroups within the data.

Given the complex, multidimensional, and interrelated nature of the variables analysed, a supervised modelling strategy would be premature or insufficient. The absence of predefined labels and the exploratory nature of the study support the application of unsupervised learning techniques.

Clustering, in particular, offers an option to identify groupings among participants based on their shared characteristics. In the subsequent modelling, body composition groupings serve as the organising lens through which we relate children's profiles to their parental context and describe the cardiorespiratory and functional traits that typify each group. This integrated child–parent perspective yields coherent profiles that reflect both physiological status and surrounding behavioural/environmental conditions, enhancing interpretability for obesity-risk assessment.

Accordingly, the next Chapter 4 applies clustering algorithms to derive meaningful participant segments and then characterises these segments across parental indices and fitness domains. These data-driven profiles will serve as a foundation for understanding heterogeneity in obesity-related risks and may help in developing targeted interventions for this specific group of participants.

4 Clustering Modelling

This chapter aims to present the methodological and analytical framework adopted for the clustering analysis. Specifically, we describe the application of hierarchical clustering over selected body composition features at time-point T1, including the hyperparameter configuration process and the interpretation strategy used to determine the adequacy of the results. The emphasis here is on methods rather than outcomes: the procedures, rationale, and analytical choices that structure the modelling approach. The empirical results derived from this framework will be presented and discussed in the following chapter.

4.1 Analytical Framework

This section introduces the analytical framework adopted in this project, the selected dataset, and the sequence of modelling steps.

To provide a structured overview of the modelling strategy, Figure 4.1 summarises the full pipeline adopted in this thesis. The analysis begins with the selected body composition features of time-point T1 obtained from the feature selection stage, which form the input for hierarchical clustering. The resulting clusters constitute the basis for subsequent steps of interpretation and evaluation.

The flow can be described in five main stages:

1. **Clustering:** Agglomerative hierarchical clustering was applied to the selected body composition variables (T1) to identify natural groupings of children. The rationale for choosing hierarchical clustering over k-means is detailed in Section 4.2.

2. **Visualisation and Cluster Profiling:** Clusters were characterised through multiple visual techniques (boxplots, stacked bar charts, radar plot), enabling the identification of both continuous and categorical differences.

3. **Cross-Domain Association Analyses:** Cluster membership was related to additional health domains (cardiorespiratory, functional, and drawing indices), as well as to parental questionnaire indices (both at time-point T1), to explore behavioural and contextual determinants.

4. **Model Interpretation:** SHAP values were employed to interpret variable contributions to the clustering process.

5. **Longitudinal Analysis:** Longitudinal analysis focused on BMI percentile trajectories across the three timepoints to evaluate BMI percentile cluster evolution over time.

This integrative flow ensures that clustering results are not considered in isolation, but are systematically connected to explanatory factors and longitudinal developments, offering a comprehensive framework for obesity risk assessment.

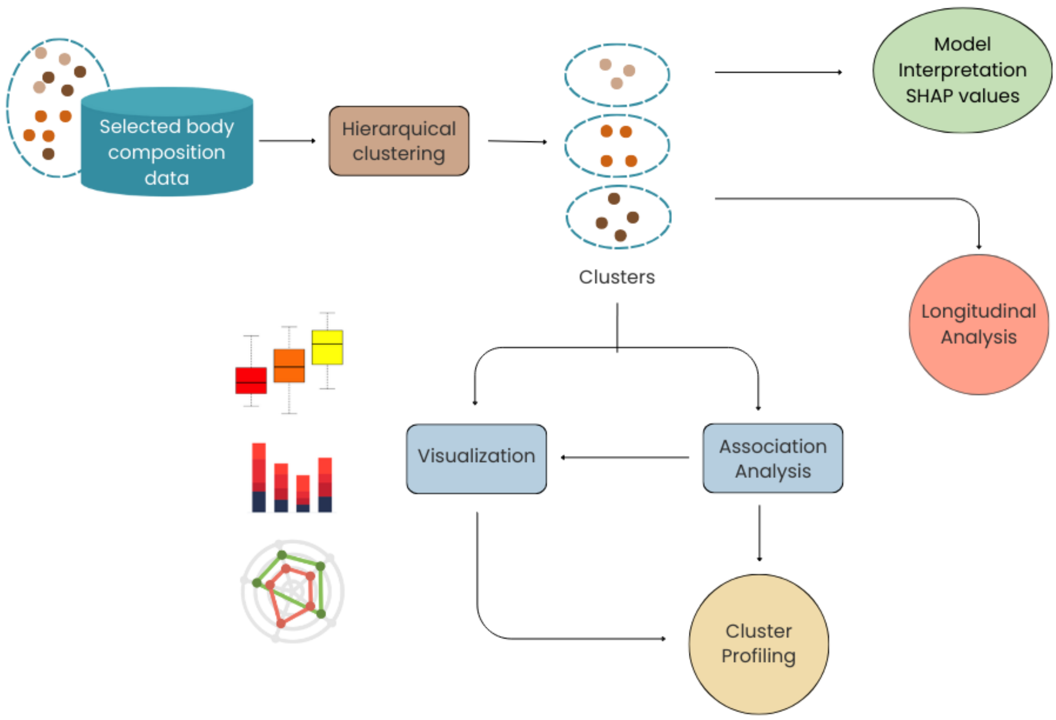


Figure 4.1: Analytical flow of the modelling approach. Hierarchical clustering is applied to body composition features at T1, followed by cluster profiling, association analyses, model interpretability, and longitudinal analysis.

4.2 Hierarchical Clustering

Hierarchical clustering is an unsupervised learning algorithm that builds nested groups of observations by either progressively splitting clusters (divisive approach) or merging them (agglomerative approach). In this study, the agglomerative strategy was applied: each observation begins as its own cluster, and at each iteration, the two most similar clusters are merged until all samples belong to a single group. This process is illustrated in Figure 4.2, where six elements (A–F) are gradually combined into larger clusters through successive merging steps. The resulting hierarchy of merges can be

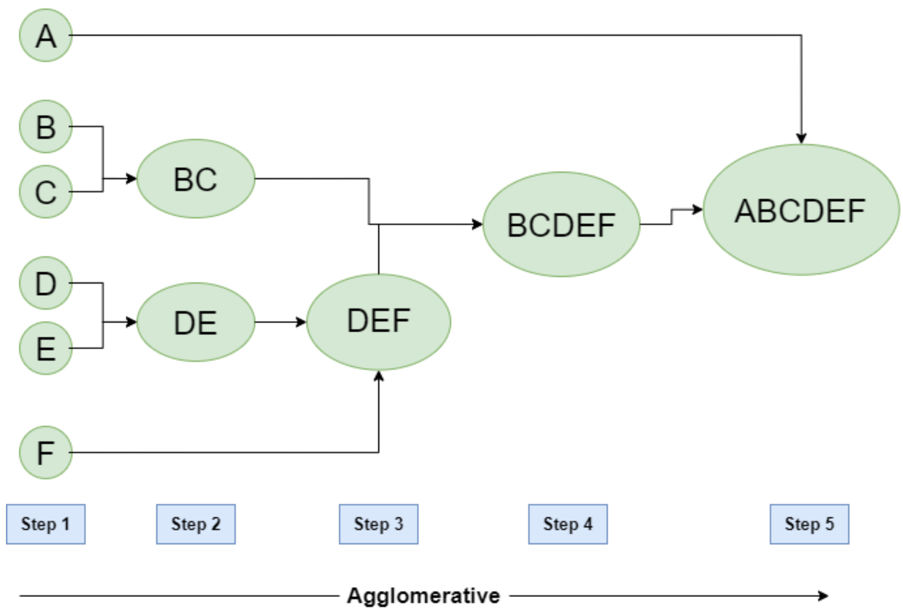


Figure 4.2: Illustration of the agglomerative hierarchical clustering process. Individual elements (A–F) start as separate clusters and are successively merged until a single cluster remains [36].

visualised through a *dendrogram*, a tree-like structure that records the order of merges and the corresponding dissimilarity distances. Dendrograms are particularly useful for visual inspection to identify natural divisions in the data and for guiding the decision of where to cut the tree at a selected level.

The Ward linkage criterion was employed, which minimises the total within-cluster variance at each step. Euclidean distance was used as the dissimilarity metric. These settings constitute the default hyperparameters of the `SciPy` function, and no additional tuning was applied in order to preserve interpretability and comparability with standard practice [37].

In this study, the clustering was applied to a dataset of 34 observations and 10 selected features from the body composition domain (as described in Section 3.2.2).

Why hierarchical instead of k-means

We use agglomerative hierarchical clustering rather than k-means because:

- **Small n (34) and reproducibility.** Hierarchical clustering is deterministic (no random starts), giving stable partitions in a small cohort; k-means can vary across runs.
- **No k upfront.** The dendrogram exposes structure at multiple resolutions and supports a transparent, data-driven cut; k-means requires choosing k in advance.
- **Cluster shape/size.** Ward's method tolerates unequal, mildly non-spherical groups; k-means favours spherical, equal-size clusters.

- **Interpretability.** Merge heights and the tree are easy to communicate to GicO multidisciplinary team.
- **Aligned objective.** Ward minimises the increase in within-cluster variance at each merge the same compactness k means targets—while retaining the hierarchy.

This choice follows the evidence summarised in the related work (Section 2.5.2), where hierarchical methods are preferred for small samples and when interpretability and stability are priorities.

4.3 Visualisation and Cluster Profiling

The profiling stage relied explicitly on *visual* diagnostics to construct an interpretable narrative for each cluster. We combined three complementary lenses to capture (i) distribution and spread, (ii) multivariate shape, and (iii) within-cluster composition: boxplots for continuous features, radar plots for multivariate profiles, and stacked bars for categorical summaries (the construction of derived categorical variables is detailed in Section 3.2.2).

Categorical summaries (stacked bars). For each derived categorical indicator (e.g., BMI category), we plot stacked bars therefore emphasise how categories distribute *inside* each cluster rather than absolute counts.

Continuous features (boxplots). Using Seaborn/Matplotlib, we drew boxplots that compare each selected body composition feature across clusters.

Radar plots (multivariate profile). The features shown on the radar had very different units and ranges, so values on some axes overshadowed the others. To make the shapes comparable, we first put the data on a logarithmic scale (which shrinks big numbers more than small ones) and then scaled each set of feature means to run from 0 to 1 within its cluster. We then drew the radar chart with evenly spaced axes so patterns could be compared at a glance.

4.4 Cross-Domain Association Analyses

It was examined how cluster membership relates to additional domains: cardiorespiratory, functional, drawing-based indices, and parental questionnaire indices. The objective was twofold: (i) an exploratory visual perspective of relationships, and (ii) formal testing with appropriate effect sizes and multiple-testing control.

Joint density (KDE) plot with BMI percentile. For each external variable X from cardiorespiratory, functional, and drawing indices, we plotted *joint* kernel-density estimates of $(X, \text{ BMI percentile})$ stratified by cluster using `seaborn.jointplot(kind = "kde", hue=Cluster)`. The default Gaussian KDE and bandwidth were used. A horizontal reference line at BMI percentile = 85 was added (axhline, dashed red) to mark a clinically relevant threshold.

Parental questionnaire indices. For parental indices, a correlation heatmap was suitable for quick screening; we computed Spearman's ρ between indices and cluster indicators. This provides an interpretable "association matrix". Also, boxplots that compare each questionnaire indices were drawn across clusters.

4.5 Model interpretability

Unsupervised methods such as hierarchical clustering do not natively tell us which variables are most responsible for the resulting groups. To bridge that gap, we used an *auxiliary explainability* approach: we trained a simple supervised model to mimic the cluster assignments using the selected features and then applied SHAP to that model. The idea is pragmatic: if a lightweight predictor can reliably map features to the discovered labels, SHAP attributions on that predictor provide a transparent, observation-level account of which features and values support each assignment.

The input to the auxiliary model is the same feature set used for clustering (selected body composition variables). The target is the cluster identifier produced by the hierarchical model. We therefore treat the auxiliary model as an explanatory proxy, not as a definitive classifier of new cases. All results are interpreted as *associations* that help characterise clusters.

We used a tree-based model (Random Forest) because it is natively supported by TreeExplainer, which yields fast and consistent SHAP values. Concretely, we fitted a RandomForestRegressor with default hyperparameters (`n_estimators=100`) to predict the cluster identifier from the selected features.

We applied `TreeExplainer` to the fitted forest and obtained local contributions for every observation–feature pair. A positive SHAP value means that, relative to the model's baseline, the feature value pushes the model output upwards (towards higher label codes); a negative value pushes it downwards. The absolute magnitude reflects the influence strength. We summarise these effects in the SHAP summary (beeswarm) plot, which ranks features by mean absolute SHAP and overlays the distribution of values.

Figure 4.3 summarises the workflow: (1) obtain cluster labels from hierarchical clustering; (2) fit an auxiliary Random Forest on the selected features to reproduce those labels; (3) compute SHAP values with `TreeExplainer`; (4) aggregate and visualise (beeswarm) to derive global rankings and cluster-wise interpretations.

Figure 4.3: Application of SHAP to the auxiliary Random Forest model. The model is trained to reproduce the discovered clusters; SHAP computes feature contributions and produces a global summary plot.

4.6 Longitudinal Analysis

The GicO study is longitudinal by design, with repeated assessments across multiple time-points (T1–T3). In this thesis, however, the primary analytic focus is the baseline (T1) cohort. This choice reflects both the phased delivery of datasets (T1 available earliest; later time-points delivered near the close of the project window) and the small sample size, which constrains the reliability of complex longitudinal models.

Given these constraints, we implemented a *minimal, descriptive* longitudinal check centred on a single relevant feature: the **BMI percentile**. Specifically, clusters were derived at T1 and then used to examine how each group's BMI percentile evolved at T2 and T3. No re-clustering was performed at later time-points and no inferential longitudinal modelling was attempted, due to the time window of this work.

This scoped approach serves two purposes: (i) to provide a transparent first look at whether T1 risk profiles show early divergence in BMI percentile; and (ii) to inform *future work* where richer longitudinal models can be applied once sample size and completeness permit.

4.7 Summary

This chapter presented the methodological framework for the modelling analyses. Hierarchical clustering served as the backbone of the unsupervised approach, with cluster profiling performed through both visualisation and interpretability techniques. Associations with the other children's datasets (cardiorespiratory, functional, and drawings) and parental questionnaires (at time-point T1) were systematically incorporated, while longitudinal modelling of BMI percentile trajectories extended the analysis beyond cross-sectional insights. The next chapter will present the results derived from these methods.

5 Results

This chapter presents the empirical findings of the clustering analysis and its downstream examinations. We first report how the clusters were identified and their distribution. We then profile clusters using body composition variables (including SHAP importance), examine cross-domain associations (cardiorespiratory, functional, and drawing indices), explore family/contextual factors from the parental questionnaire.

Finally, given the longitudinal design of the GicO study, we extended the analysis to examine how children's body composition evolved across the three assessment moments (T1, T2, and T3), specifically, how BMI percentile evolved.

5.1 Cluster Identification

After applying hierarchical agglomerative clustering to the selected body composition variables at time-point T1, we obtained the dendrogram shown in Figure 5.1. The structure of merges was carefully inspected to determine the optimal cutting level.

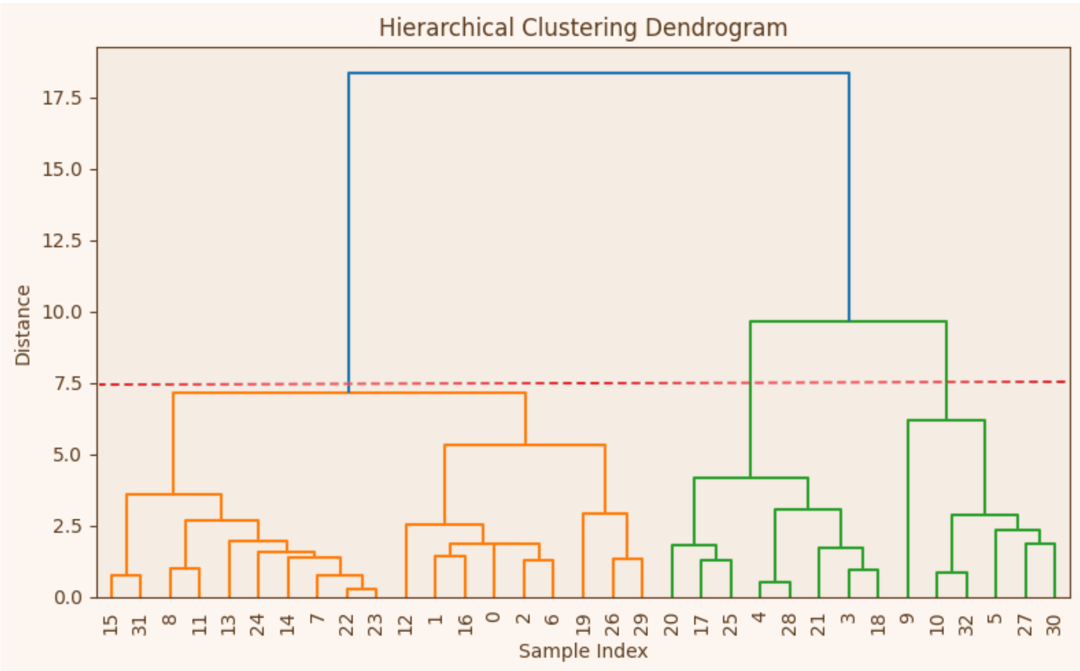


Figure 5.1: Hierarchical clustering dendrogram using selected body composition features at T1.

Following the standard practice established by Everitt et al. [38], namely, selecting the cut level at the largest increase in fusion distance, we examined the linkage heights and identified a pronounced jump at approximately **7.5**. Cutting the dendrogram at this height yields a solution of three clusters that balances interpretability with intracluster cluster homogeneity. In contrast, higher cuts (e.g., 12.5) would over-aggregate distinct profiles into overly broad categories, whereas lower cuts would fragment the cohort into less meaningful subdivisions.

To aid interpretation, cluster assignments were projected onto the first two principal components (Figure 5.2), allowing visual inspection of group separation. The first component explained **72.34%** of the variance and the second **19.43%**, for a combined **91.78%** of total variance captured by the two-dimensional projection. The corresponding sample counts per cluster are reported in Figure 5.3. The overall silhouette score was **0.34**, which indicates moderate separation. This is an expected outcome given the heterogeneity of body composition data and the relatively small sample size.

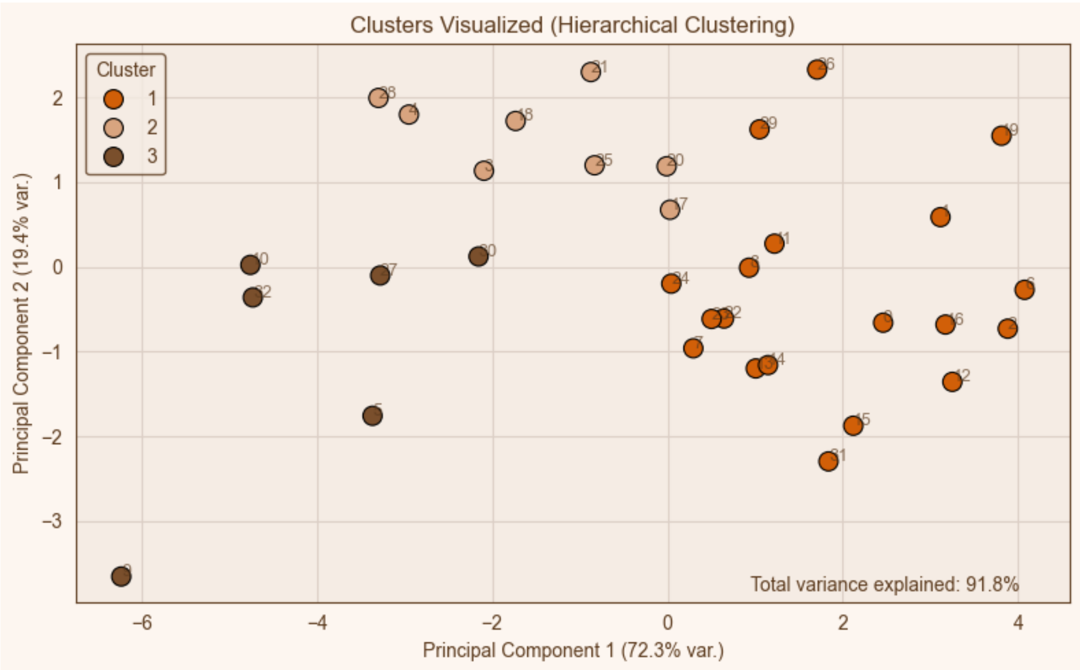


Figure 5.2: Projection of cluster assignments on the first two principal components.

5.2 Cluster Profiling

This section characterises the three clusters obtained in the previous Section 5.1 to translate the results of the unsupervised solution into interpretable profiles. We summarise the children´s body composition patterns using: (i) a radar profile per cluster (ii) distributions of continuous variables stratified by cluster, and (iii) categorical summaries of body composition categories by cluster. The objective is the interpretation of the groups.

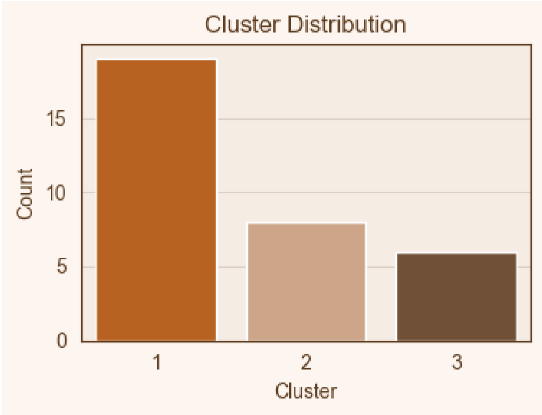


Figure 5.3: Sample counts per cluster.

5.2.1 Radar profile

Clusters are displayed in the radar plot (Figure 5.4). **Cluster 1** shows comparatively higher values on lean mass and cellular water indicators (Protein, SLM, SMM, BCM, ICW), suggesting a more favourable body composition pattern. **Cluster 3** exhibits the lowest values across most features, indicative of a comparatively poorer body composition pattern. **Cluster 2** exists between these extremes.

5.2.2 Distribution of continuous features

Boxplots stratified by cluster for the selected continuous body-composition variables are shown in Figure 5.5. For most markers, values progress consistently from lower to higher clusters.

- **Lean-mass and cell–mass markers.** ICW, SLM, SMM, BCM, FFMI, and BMR show the same pattern: *Cluster 1* exhibits the highest central values with relatively compact IQRs; *Cluster 3* shows the lowest central values; *Cluster 2* lies consistently between, with wider dispersion for some variables. This confirms that separation is primarily driven by lean tissue and cellular hydration.
- **Hydration ratio.** TBW/FFM presents smaller differences between cluster and overlapping IQRs, central values present this order (Cluster 1 and Cluster 2 $>$ Cluster 3).
- **Phase angle.** The 50 kHz-whole body phase angle mirrors the lean-mass pattern (Cluster 1 highest, Cluster 3 lowest), reinforcing differences in cellular integrity and health.
- **BMI percentile.** A pronounced shift is observed: Cluster 1 centres at higher BMI percentiles, Cluster 3 at lower percentiles, and Cluster 2 covers a wide range with high variability. This gradient aligns with the distribution seen in other composition markers.

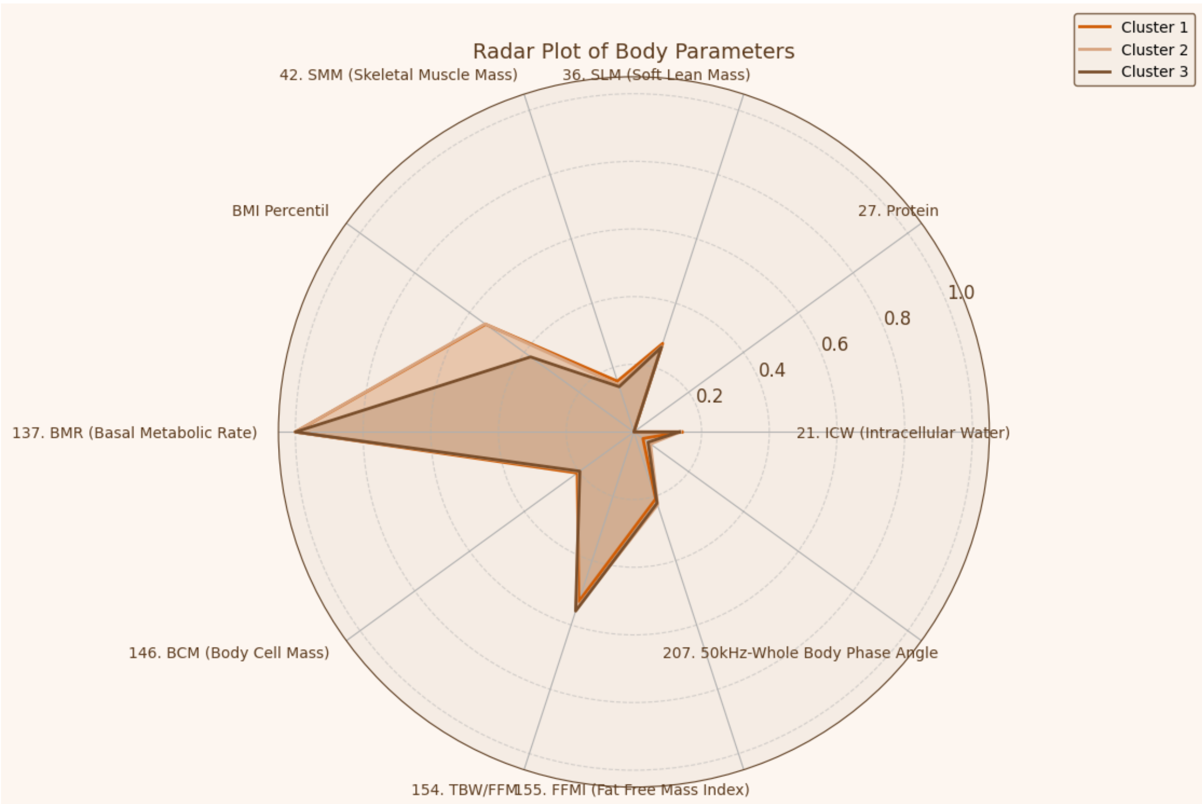


Figure 5.4: Radar plot of selected body composition features by cluster. Cluster 1 shows higher lean-mass/cell-mass markers; Cluster 3 shows comparatively lower values; Cluster 2 is intermediate.

- **Age.** Age shows no systematic separation across clusters (similar central values and overlapping interquartile range), suggesting that age is unlikely to be a confounder of the clustering solution.

Overall, the boxplots corroborate the radar summary: clusters differ most strongly in lean-mass and cellular-water features, with Cluster 2 displaying greater heterogeneity. These distributional contrasts provide visual evidence of the separation between the clusters and support the interpretability of the groups.

5.2.3 Categorical summaries

Category-distribution visualisations stratified by cluster for the derived categorical variables (Chapter 3, Section 3.4) are shown in Figures 5.6–5.8. Three consistent patterns emerge:

1. **Lean and cell–mass categories** (e.g., SLM, SMM, FFM, BCM, Protein, Minerals). The Cluster 1 shows a larger share of favourable categories (HIGH/NORMAL), whereas the Cluster 3 concentrates more *LOW* categories. The Cluster 2 exhibits mixed composition with proportions between the two extremes.
2. **Water compartments** (TBW, ICW, ECW). Cluster 1 is skewed towards NORMAL/HIGH TBW and ICW, while Cluster 3 shows a visible shift towards LOW. ECW differences are more modest but remain directionally consistent with the overall pattern.
3. **Adiposity-oriented categories** (PBF, WHR, Obesity Degree, BMI category). Cluster 3 displays a higher prevalence of *less favourable* categories (e.g., OVERWEIGHT/OBESITY or HIGH adiposity), whereas Cluster 1 concentrates NORMAL/LOW categories; Cluster 2 is intermediate with broader spreads.

These categorical distributions reinforce the narrative established by the continuous summaries (Section 5.2.2): separation across clusters is mainly driven by lean-mass and cellular-water status, with adiposity categories aligning accordingly.

In summary, the observed patterns define three distinct profiles (Table 5.1): (i) a normal (favourable composition, lower risk) — Cluster 1, (ii) a normal (average composition) — Cluster 2, and (iii) a higher-risk profile (less healthy composition) — Cluster 3. In other words, both Cluster 1 and Cluster 2 are “normal”, but with distinct characteristics.

Table 5.1: Summary of the profiles derived from body composition clusters.

Cluster	**Profile label**
Cluster 1	**Normal** (favourable composition)
Cluster 2	**Normal** (average composition)
Cluster 3	**Higher-risk** (less healthy composition)

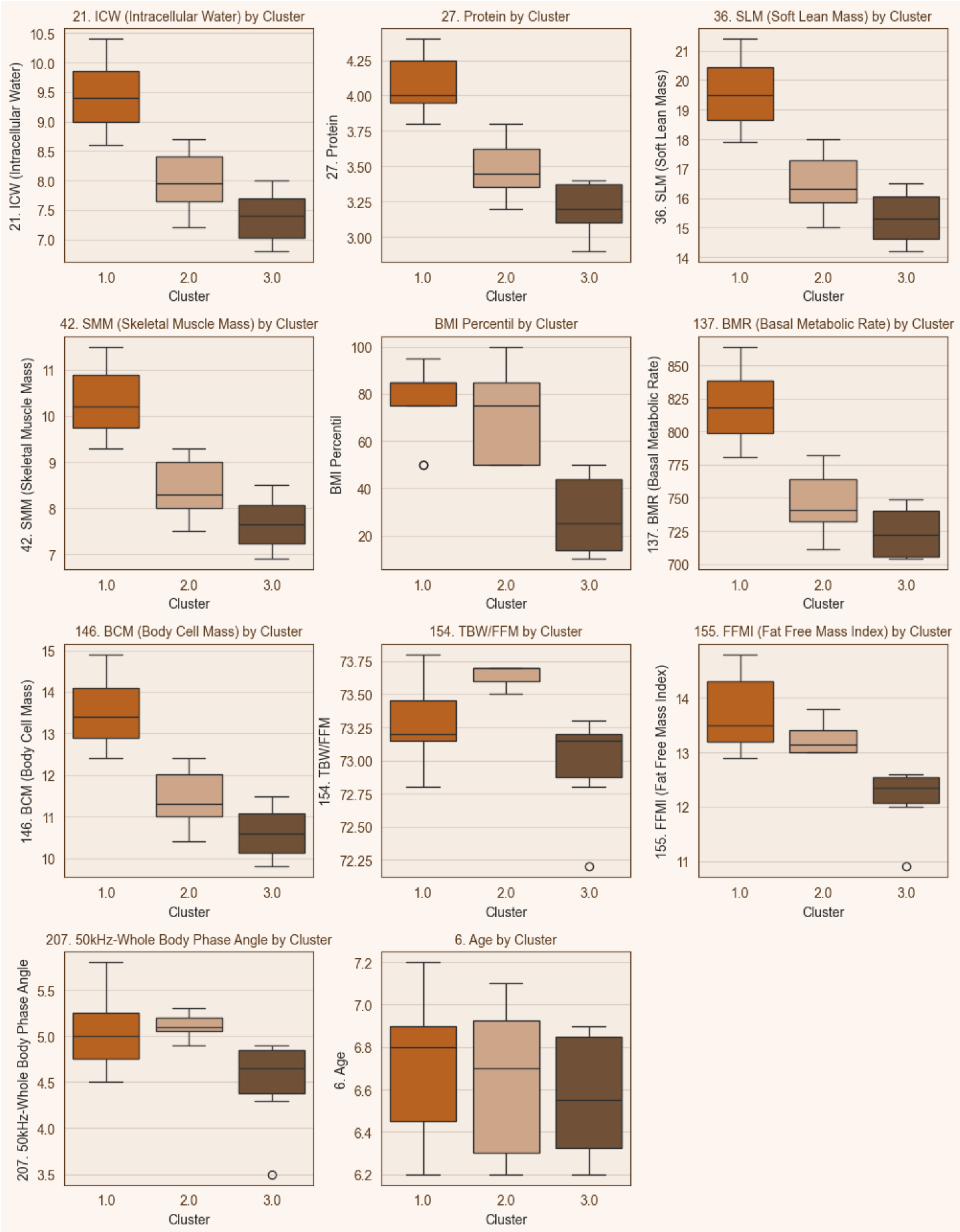
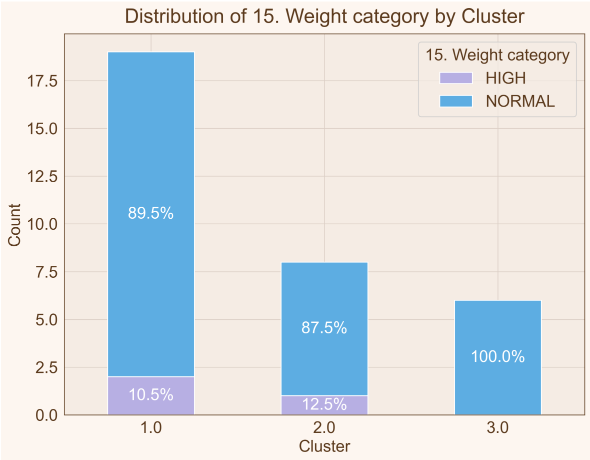
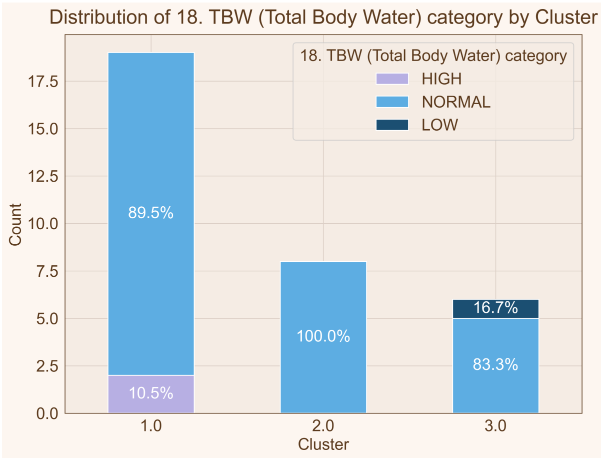


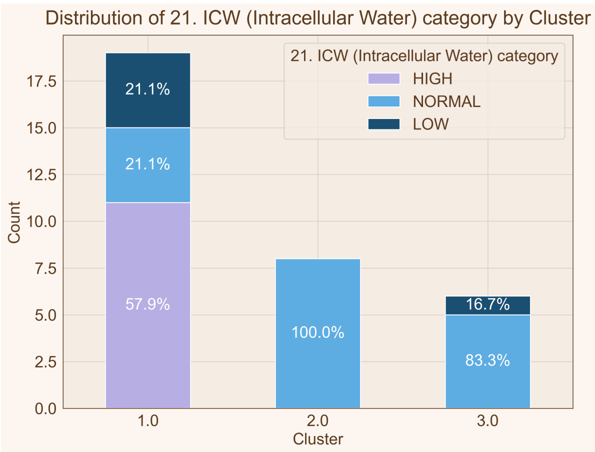
Figure 5.5: Boxplots for continuous body composition variables. central values and IQRs show a clear gradient across lean–mass and cellular–water markers (Cluster 1 highest, Cluster 3 lowest).



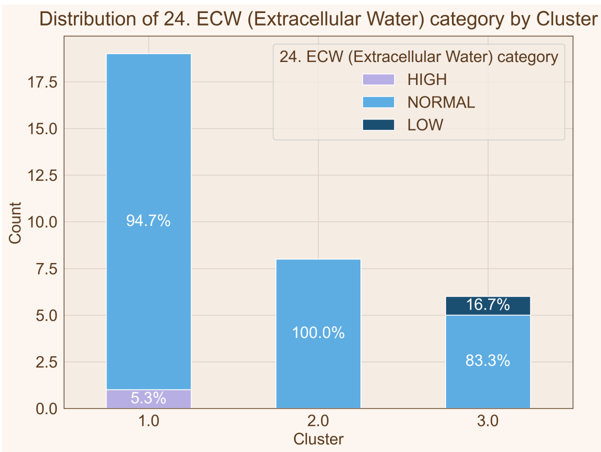
(a) Weight.



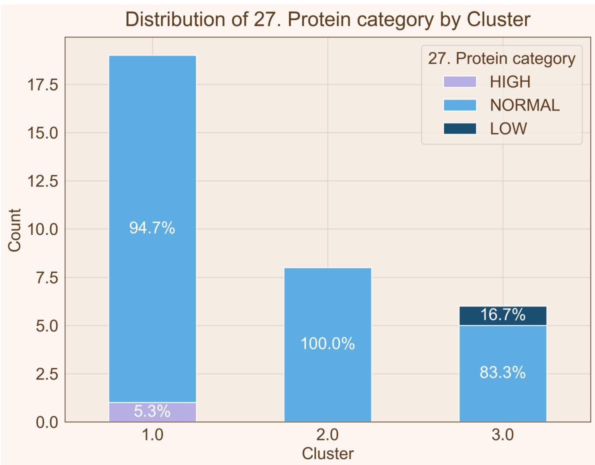
(b) TBW.



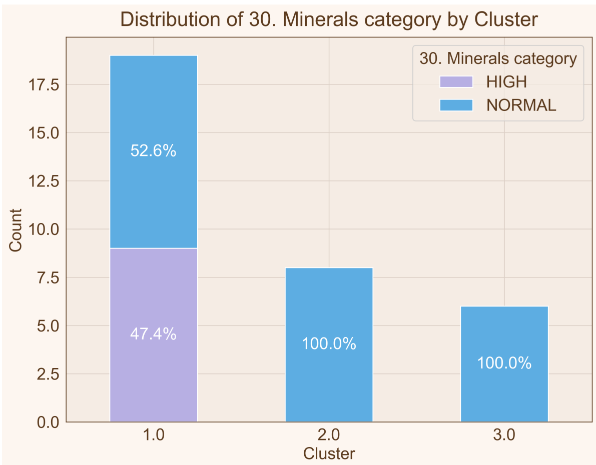
(c) ICW.



(d) ECW.

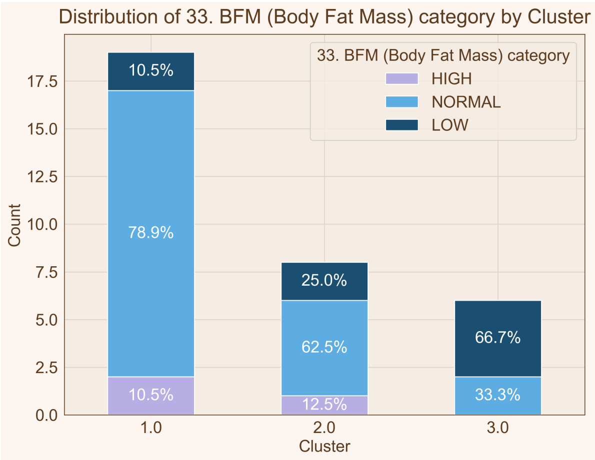


(e) Protein.

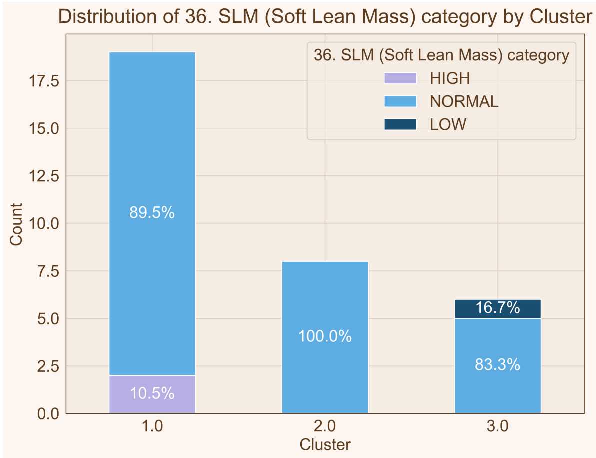


(f) Minerals.

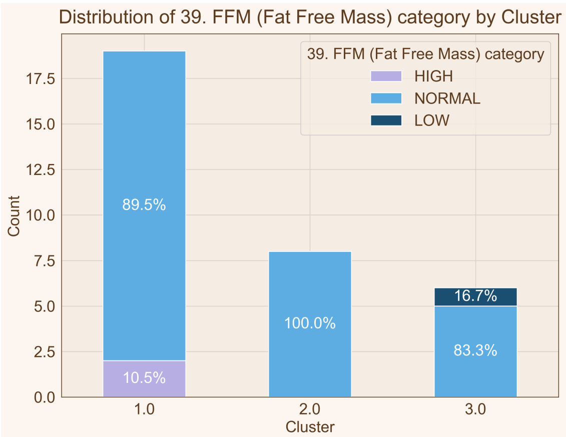
Figure 5.6: Distributions of body composition categories by cluster (Part 1/3).



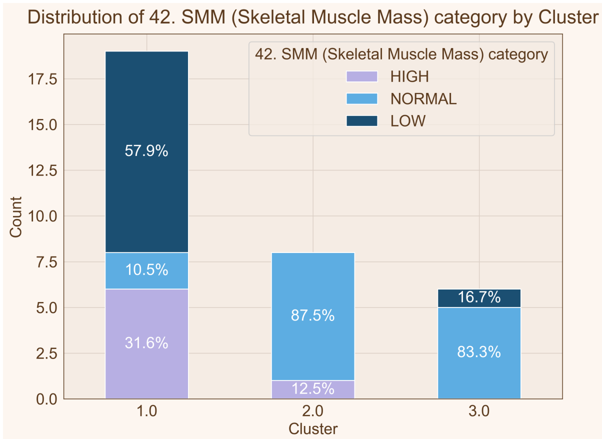
(a) BFM.



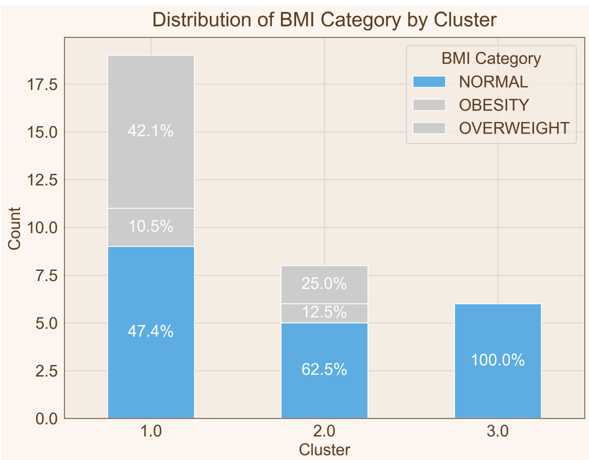
(b) SLM.



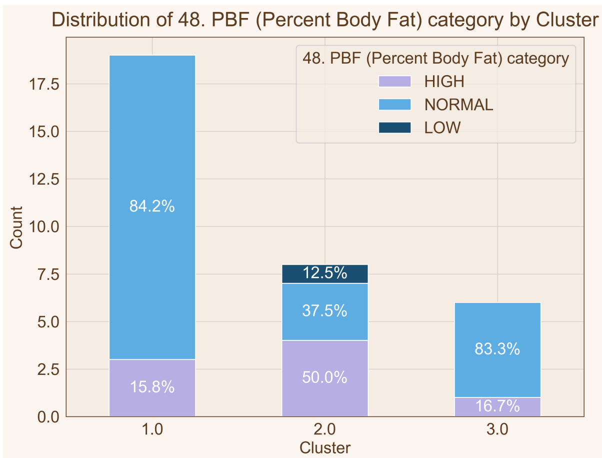
(c) FFM.



(d) SMM.

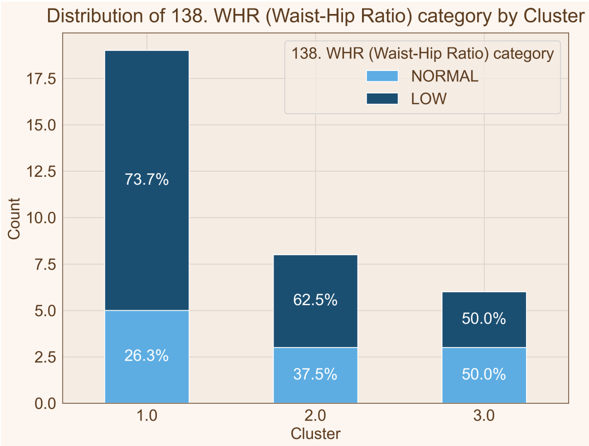


(e) BMI.

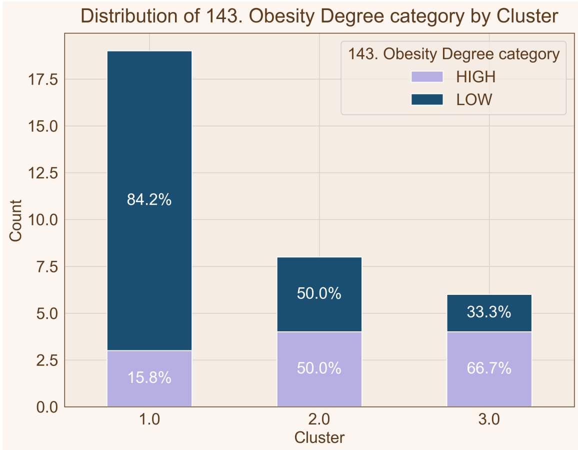


(f) PBF.

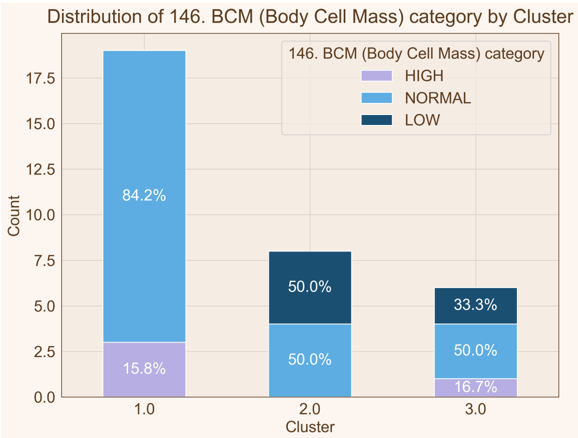
Figure 5.7: Distributions of body composition categories by cluster (Part 2/3).



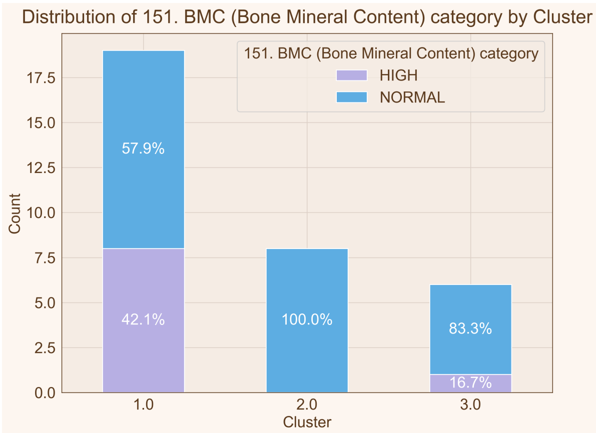
(a) WHR.



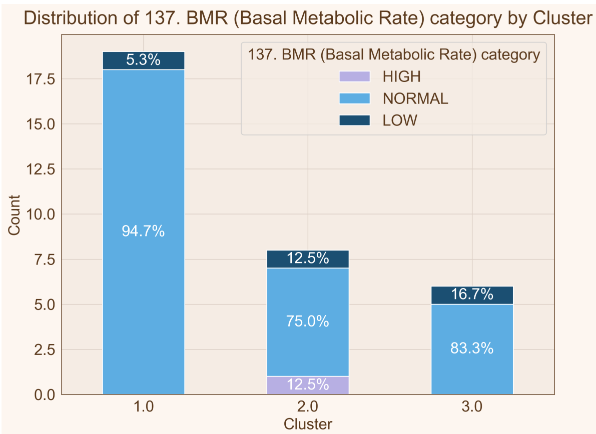
(b) Obesity degree.



(c) BCM.



(d) BMC.



(e) BMR.

Figure 5.8: Distributions of body composition categories by cluster (Part 3/3).

5.3 Cross-Domain Associations

We examined how the body composition clusters relate to other measurement domains: cardiorespiratory fitness, functional performance, family context (parental questionnaire indices), and drawing-based perceptions. For cardiorespiratory and functional variables, we used joint kernel density estimate (KDE) plots of each domain variable X against BMI percentile, stratified by cluster. For the family context, we summarised indices in a heatmap, and for children´s perceptions through drawings, we used bar plots. BMI percentile is used as the reference axis for bivariate plots because it is the study's primary feature agreed with GicO domain experts, and the body composition feature selection was explicitly oriented to this target; moreover, being standardised by age and sex, it offers an interpretable common scale.

5.3.1 Cardiorespiratory domain

Figure 5.9 shows joint KDE plots stratified by cluster of BMI percentile against the four cardiorespiratory metrics at time-point T1. Across subplots, the density ridges display a clear *inverse pattern*: better cardiorespiratory performance aligns with lower BMI percentiles, whereas higher BMI percentiles concentrate where performance is weaker. The horizontal dashed line marks the 85^{th} BMI percentile as a pragmatic reference for overweight classification based on the Portuguese Directorate-General of Health [23] BMI percentile classification.

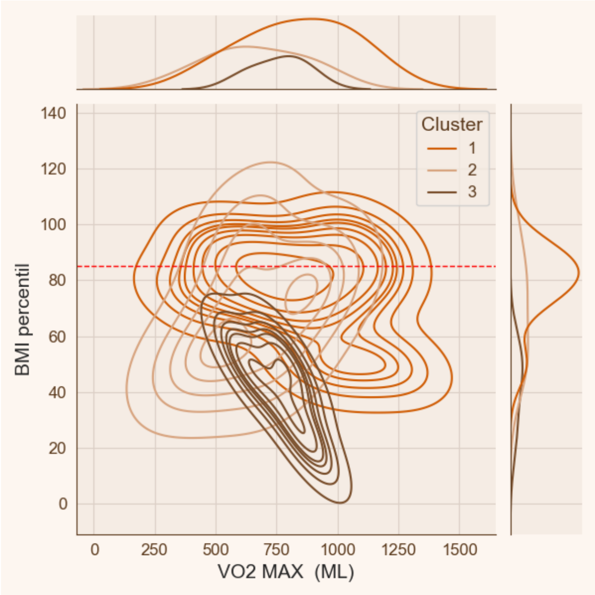
VO_2 (mL) and VO_2 (mL/kg/min). Both absolute and normalised oxygen uptake show inclined contours: higher VO_2 values match with lower BMI percentiles, with the healthier composition cluster concentrating density below the 85^{th} percentile and at higher VO_2. The less favourable composition cluster places more mass above the 85^{th} percentile with attenuated VO_2. The normalised plot (mL/kg/min) slightly accentuates this separation, as expected when accounting for body size.

Power (W). A similar trend emerges for cycling power: higher power regions are primarily located at lower BMI percentiles, while densities at higher percentiles are associated with lower power. The intermediate cluster overlaps both regimes, reflecting heterogeneity within this group.

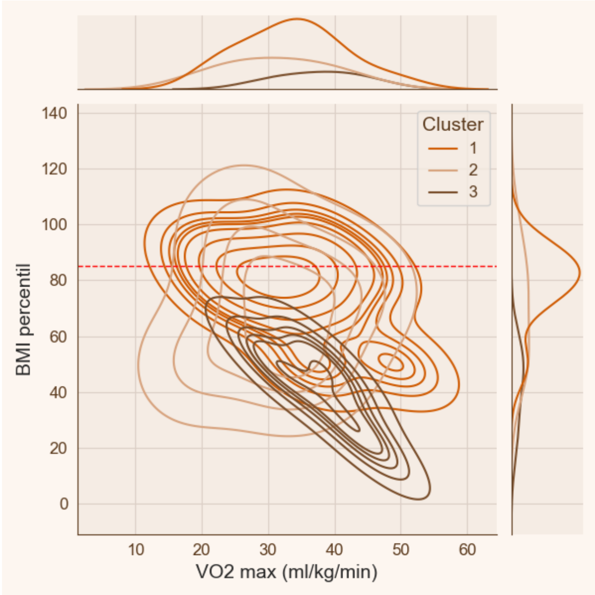
Time to exhaustion (s). Longer times cluster with lower BMI percentiles; shorter times are more prevalent at higher percentiles. The distribution again places the healthier cluster mainly below the 85^{th} percentile with longer times, and the less favourable cluster above that line with shorter times.

Taken together, the four subplots consistently suggest that *cardiorespiratory fitness is in-*

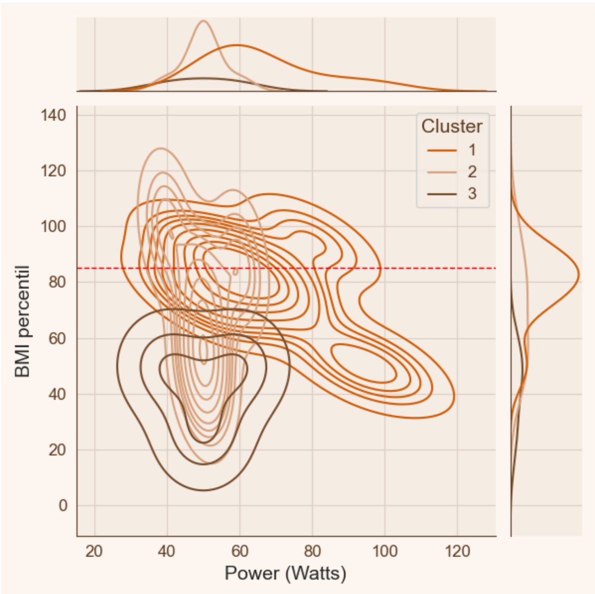
versely related to BMI percentile: higher fitness (greater VO_2, power, and time) corresponds to lower BMI percentiles, with clusters separating accordingly.



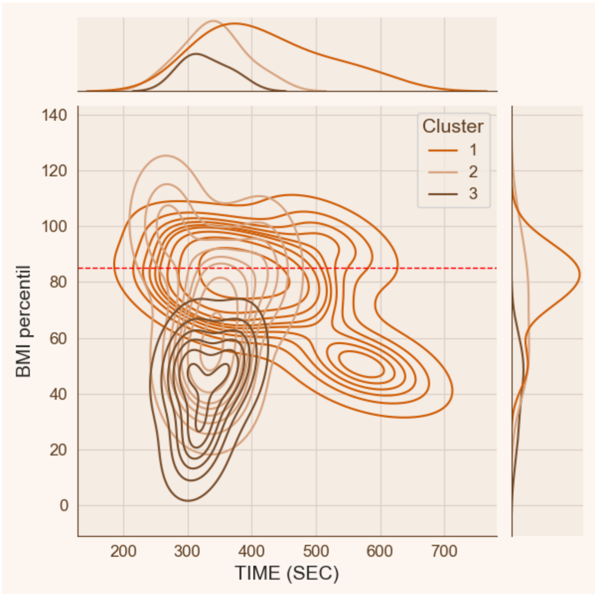
(a) VO_2 max (mL) vs. BMI percentile.



(b) VO_2 max (mL/kg/min) vs. BMI percentile.



(c) Power (W) vs. BMI percentile.



(d) Time to exhaustion (s) vs. BMI percentile.

Figure 5.9: Joint KDEs by cluster for cardiorespiratory variables against BMI percentile (T1).

5.3.2 Functional domain

Figures 5.10 and 5.11 show joint KDE plots stratified by cluster between BMI percentile and the five functional metrics at T1. Across subplots, the contours generally fall from right to left, indicating that *better functional performance tends to align with lower BMI percentiles*. The 85^{th} BMI percentile is also included as a visual reference for overweight classification as referenced in the previous section.

Seat-and-reach (flexibility). Contours for the three clusters overlap substantially, with only a mild tendency for lower BMI percentiles to coincide with greater flexibility. Any separation is modest, suggesting flexibility contributes less to differences between clusters than strength or power measures.

Handgrip strength (right/left). For both hands, higher grip values concentrate more frequently below the 85th BMI percentile, whereas lower grip values are common at higher percentiles. The cluster located predominantly at lower BMI percentiles shows broader support towards higher grip, while the cluster with higher BMI percentiles concentrates density at lower–mid grip ranges. This pattern is consistent with the notion that muscular strength relates inversely to BMI percentile in this cohort.

Jump performance (squat jump and countermovement jump). Jump metrics display the clearest trends: higher jump heights are associated with lower BMI percentiles, with contours for the lower-BMI cluster extending further into the high-performance region. In the other hand, densities above the 85th percentile are skewed towards lower jump heights. The intermediate cluster spans both regions, reflecting heterogeneity within this group.

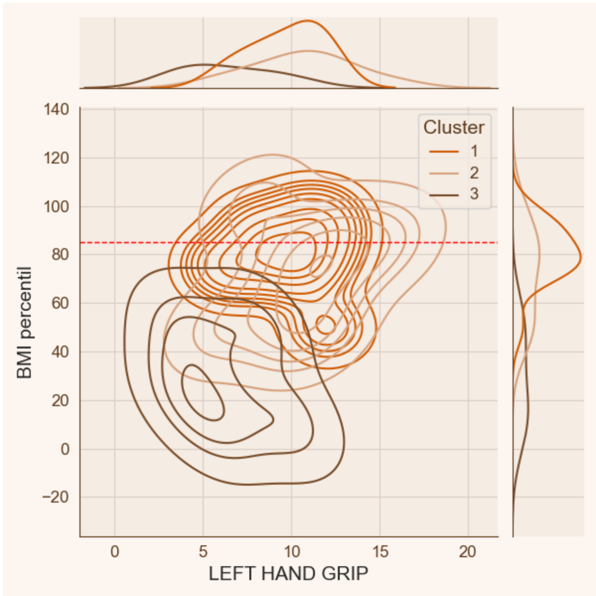
Taken together, the functional domain shows a *negative association* between BMI percentile and strength/power performance, strongest for squat and countermovement jumps, moderate for handgrip, and weakest for seat-and-reach. These visual results complement the body composition profiling: clusters characterised by more favourable composition also tend to occupy more favourable functional regions.

5.3.3 Family context

To make the connection between the child and their home environment easier to understand, we begin by asking a simple question: *do families who report different behaviours and contexts also tend to appear in different body composition clusters?* To explore this, we plot a heatmap that relates the cluster to parental questionnaire indices (environmental and behavioural domains), using Spearman's ρ to capture correlation values. The goal is to discover whether the family context, reported by parents through the questionnaires, shows correlations with the already identified clusters, i.e., whether family factors might influence or be associated with children's body composition profiles. For completeness, we also present boxplots stratified by cluster for each parental index to visualise distributional shifts.

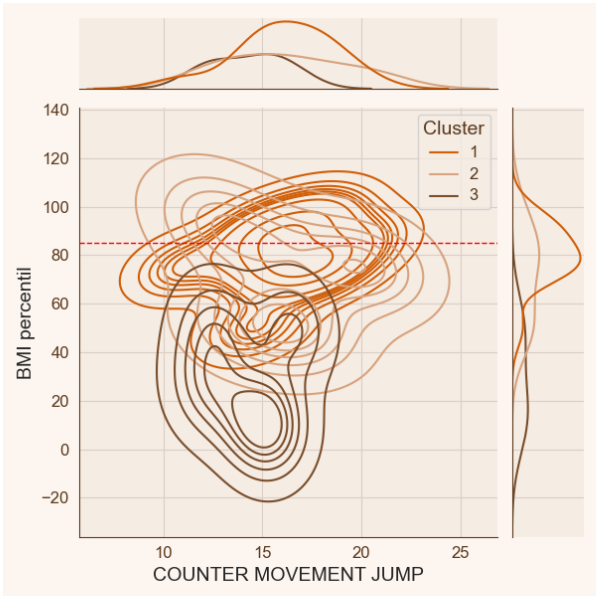
The heatmap in Figure 5.12 recovers the expected structure among child variables (e.g., BMI percentile is positively related to obesity degree and inversely to FFMI). Alongside the *cluster* indicator, we also included a small set of body composition markers to aid interpretation (e.g., FFMI, ICW, phase angle), helping to identify whether any ap-

(a) Seat and reach (cm) vs. BMI percentile.



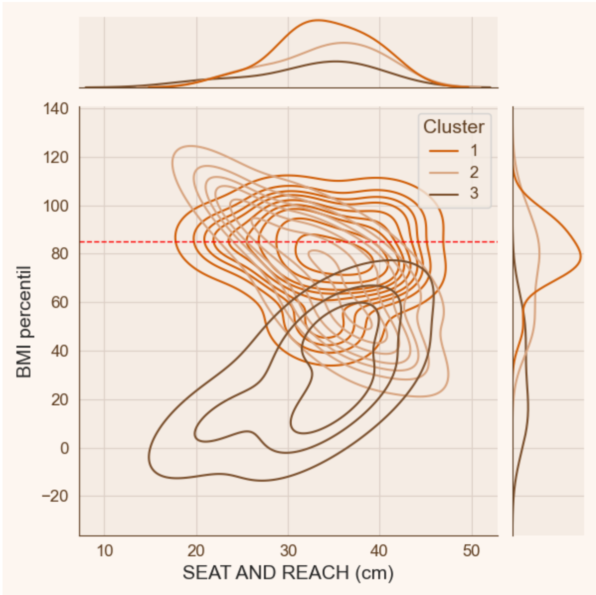
(b) Right hand grip (kg) vs. BMI percentile.

(c) Left hand grip (kg) vs. BMI percentile.



(d) Squat jump (cm) vs. BMI percentile.

Figure 5.10: Joint KDEs by cluster for functional metrics against BMI percentile (T1) (Part 1/2)



(a) Countermovement jump (cm) vs. BMI percentile.

Figure 5.11: Joint KDEs by cluster for functional metrics against BMI percentile (T1) (Part 2/2)

parent associations with parental indices might reflect underlying composition itself. Correlations between the *cluster* indicator and parental indices are generally modest, suggesting heterogeneous family contexts across clusters; nonetheless, any consistent departures (positive or negative) highlight domains where family environment and behaviour may relate to children's composition profiles and merit closer inspection.

Figures 5.13 summarise distributional contrasts:

- **PAI (Physical Activity Index).** Cluster 2 exhibits the highest median PAI; Clusters 1 and 3 are lower, with Cluster 1 showing wider dispersion.
- **STI (Screen Time–Sedentary).** Central values are highest in Cluster 3, lowest in Cluster 2, with Cluster 1 intermediate.
- **DQI (Diet Quality).** Cluster 1 has the highest central tendency; Clusters 2 and 3 are lower and comparatively similar.
- **SFL (Sustainable Food Literacy).** Cluster 3 is highest, Cluster 1 moderate, Cluster 2 lowest.
- **GAI (Green Access).** Cluster 3 shows the most favourable levels; Cluster 1 is moderate with larger variability; Cluster 2 is lower.
- **SQI (Sleep Quality).** Cluster 1 tends to higher values, Cluster 2 lower, Cluster 3 between them.
- **Stress.** Cluster 3 presents the highest central value stress; Cluster 2 the lowest; Cluster 1 is intermediate with wider spread.

- **Socio-economic status (SES).** Cluster 3 has the highest central value SES; Cluster 2 is lower; Cluster 1 is moderate but more dispersed.

Although not uniform across all indices, tendencies align with expectations: less favourable parental contexts (higher Screen (sedentary) time , higher stress) occur more often in Cluster 3, whereas more favourable contexts (better diet quality, better sleep) are observed in Cluster 1. Cluster 2 frequently occupies an intermediate position and shows the highest parental physical activity.

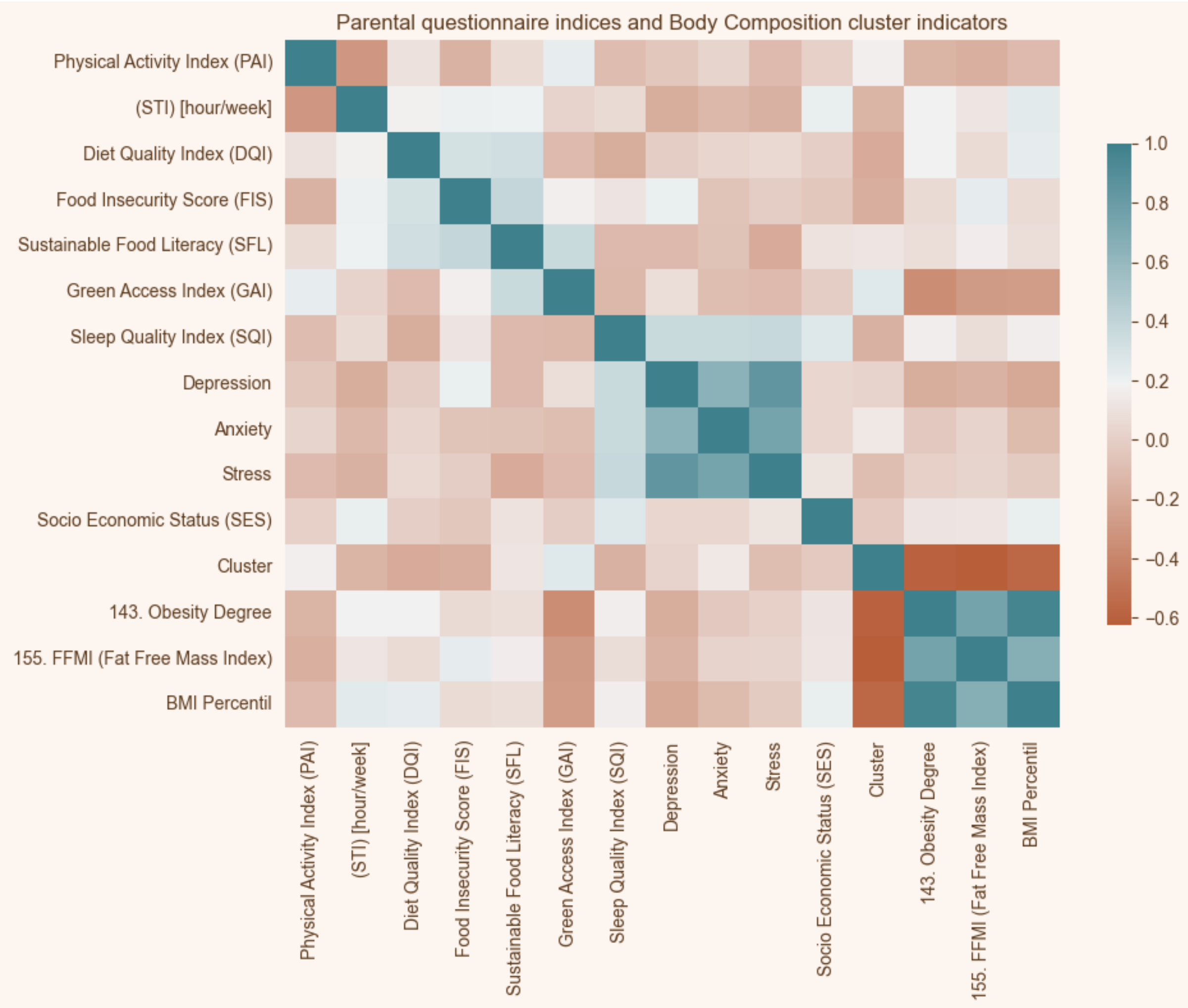


Figure 5.12: Correlation matrix (Spearman's ρ) among parental/contextual indices, the cluster indicator, and selected body composition metrics.

5.4 Children's Perceptions through Drawings

This section explores qualitative indicators derived from children's drawings as indirect indicators of perception, rather than behavioural measurements. The indicators summarise whether drawings contained ultra-processed foods, counts of food- and activity-related elements, and evidence of sedentary activity. Given the small sample

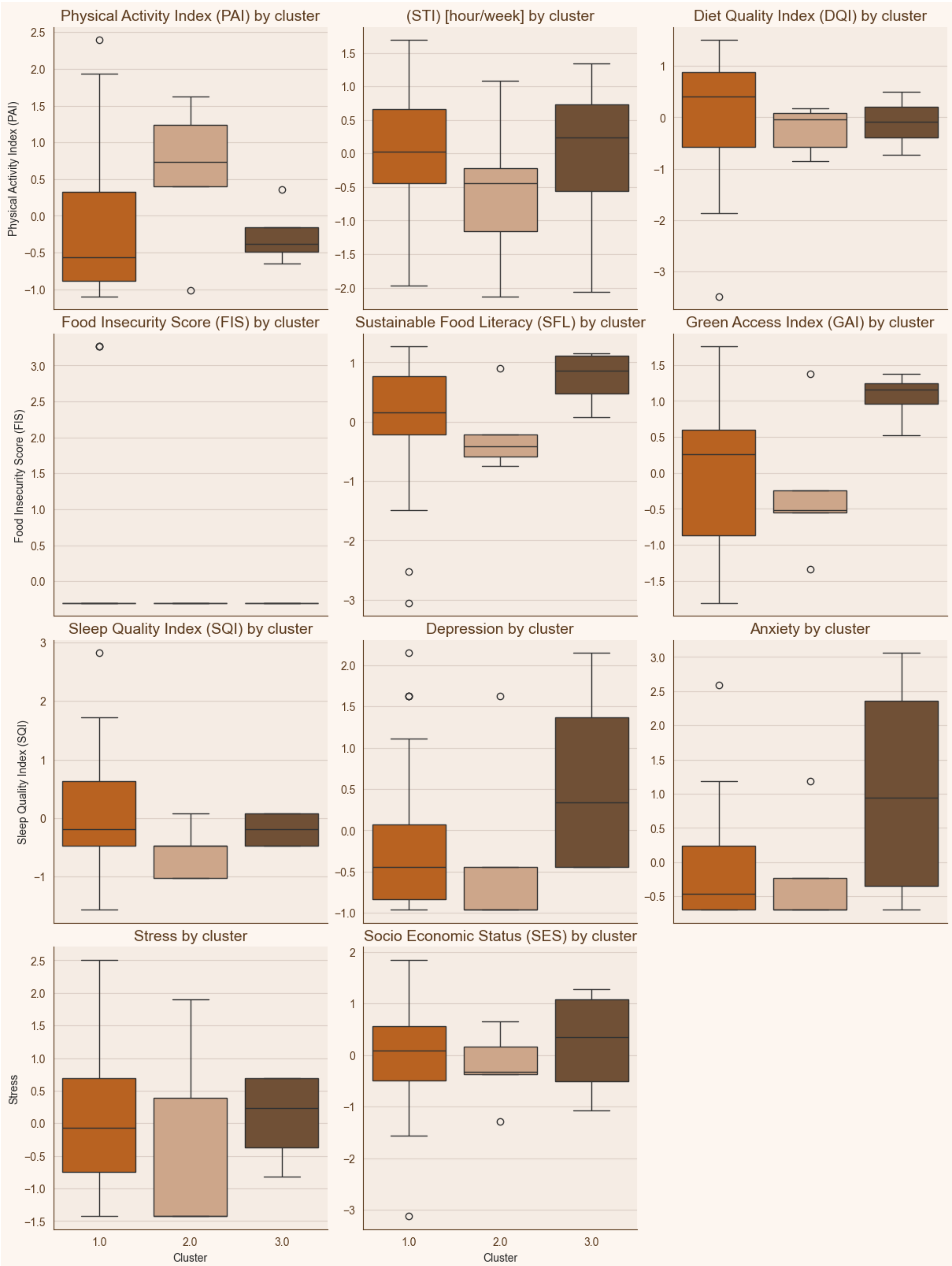


Figure 5.13: Boxplots for behavioural, dietary, sleep, and mental-health indices.

and the subjective nature of coding, the results are descriptive.

Ultra-processed foods. Figure 5.14 suggests comparable frequencies across clusters with a slight tendency for Cluster 1 to include ultra-processed items more often. As drawings capture salience rather than intake, this does not imply higher consumption but may reflect brand recall or visual prominence.

Activity indicator count. As shown in Figure 5.15, most children across clusters drew at least one activity element. Cluster 1 concentrates around a single activity item, Cluster 2 shows a similar but slightly broader spread (including some with two items), and Cluster 3 presents fewer activity depictions overall.

Food indicator count. Figure 5.16 indicates that Cluster 2 tends to include a greater number of distinct food elements (longer right tail), while Cluster 1 centres around one to two items, and Cluster 3 shows fewer food depictions.

Sedentary activity. Depictions of sedentary behaviours are generally infrequent (Figure 5.17); most drawings fall into the lower category across clusters, with occasional higher values in Cluster 2.

Overall, the drawings indicate modest cluster differences in what children choose to represent: Cluster 2 contains more food-related elements., Cluster 1 depicts activity elements most consistently, and Cluster 3 tends to include fewer food/activity indicators. Because these indicators capture perceptions and visibility, they should be read as complementary context to the objective body composition and fitness findings, not as evidence of actual behaviours.

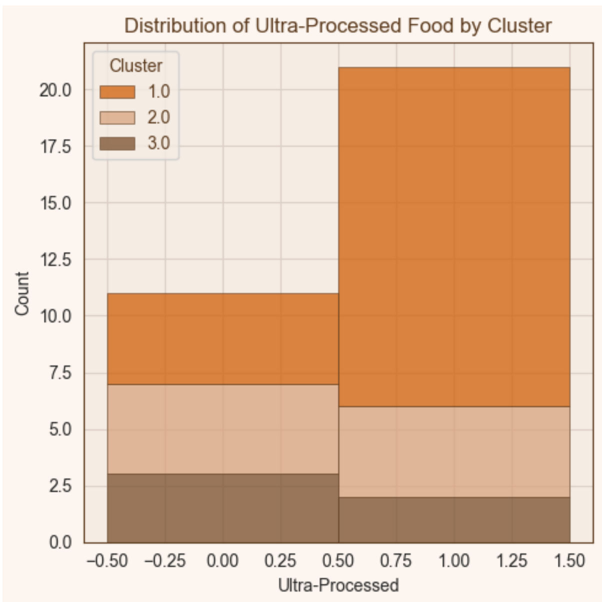


Figure 5.14: Distribution of drawings that include ultra-processed foods by cluster.

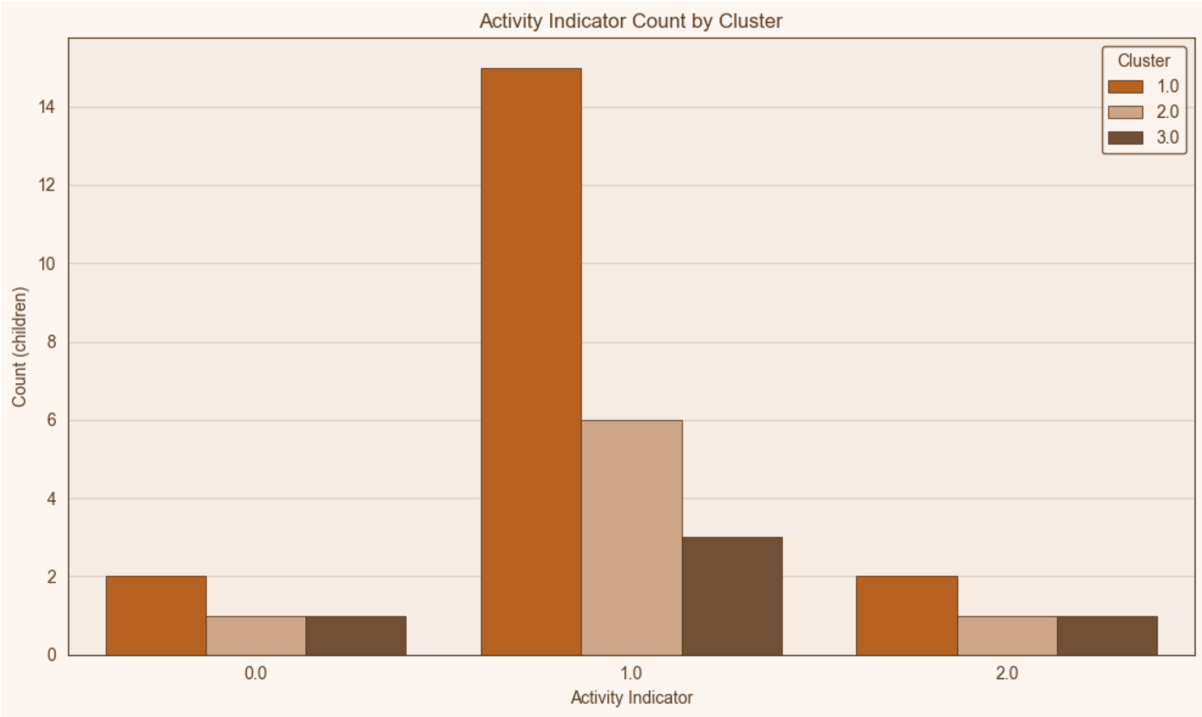


Figure 5.15: Activity indicator count by cluster.

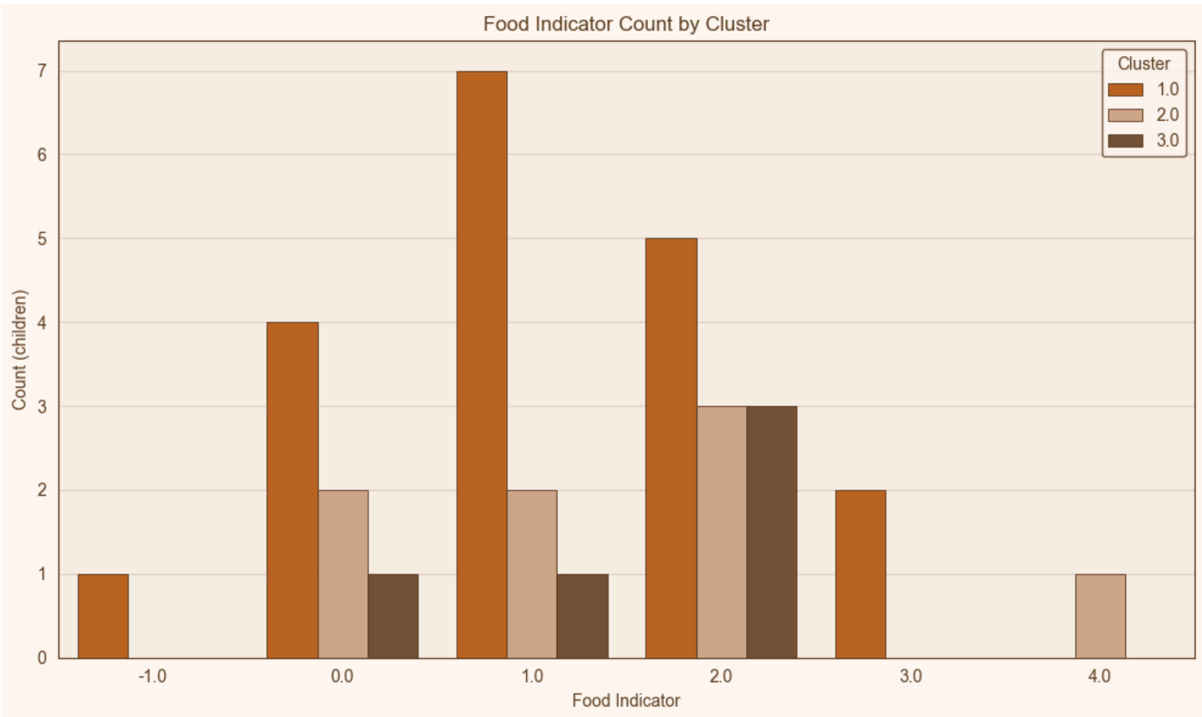


Figure 5.16: Food indicator count by cluster (number of distinct food elements depicted).

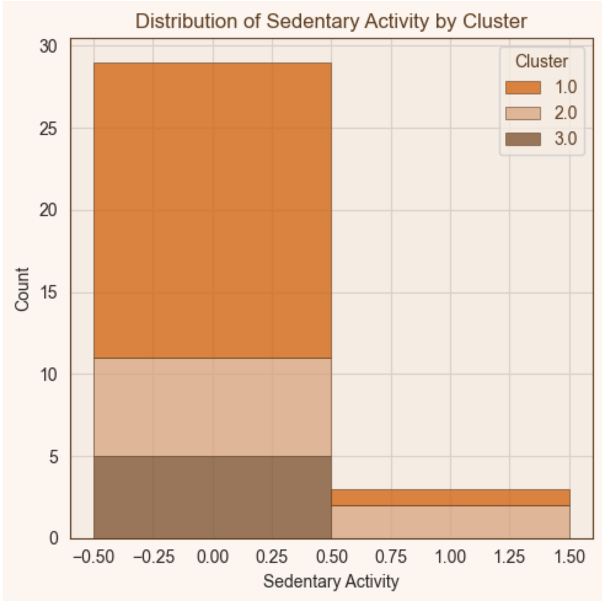


Figure 5.17: Distribution of sedentary activity depictions by cluster.

5.5 Model interpretability - SHAP values

To understand which body composition variables most strongly align with the discovered clusters, we fitted an *auxiliary* random forest to predict cluster labels from the selected features and computed SHAP values (Chapter 4). The global summary (Figure 5.18) indicates that **lean mass and cellular–water markers dominate the ranking**: Protein, SLM, BMR, FFMI, TBW/FFM, BCM, ICW, SMM, and Phase Angle appear among the most influential variables. Qualitatively, *lower* values in these markers tend to push predictions towards the higher cluster index, consistent with a less healthy body composition pattern.

5.6 Longitudinal Results

Finally, given the longitudinal design of the GicO study, BMI percentile trajectories were examined across the three assessment moments. Although the design includes three time points, participant attrition reduced sample sizes after T1. Consequently, longitudinal analysis was limited to participants with complete follow-up data. Mean cluster trajectories were plotted with 95% confidence intervals, enabling an evaluation of whether baseline cluster membership predicted differential developmental patterns over time. This analysis was crucial to assess the stability and evolution of obesity risk profiles.

Confidence bands are the shaded areas around each line showing our uncertainty about the *cluster average* at each time point. If we could repeat the study many times under the same conditions, about 95% of those bands would cover the true cluster

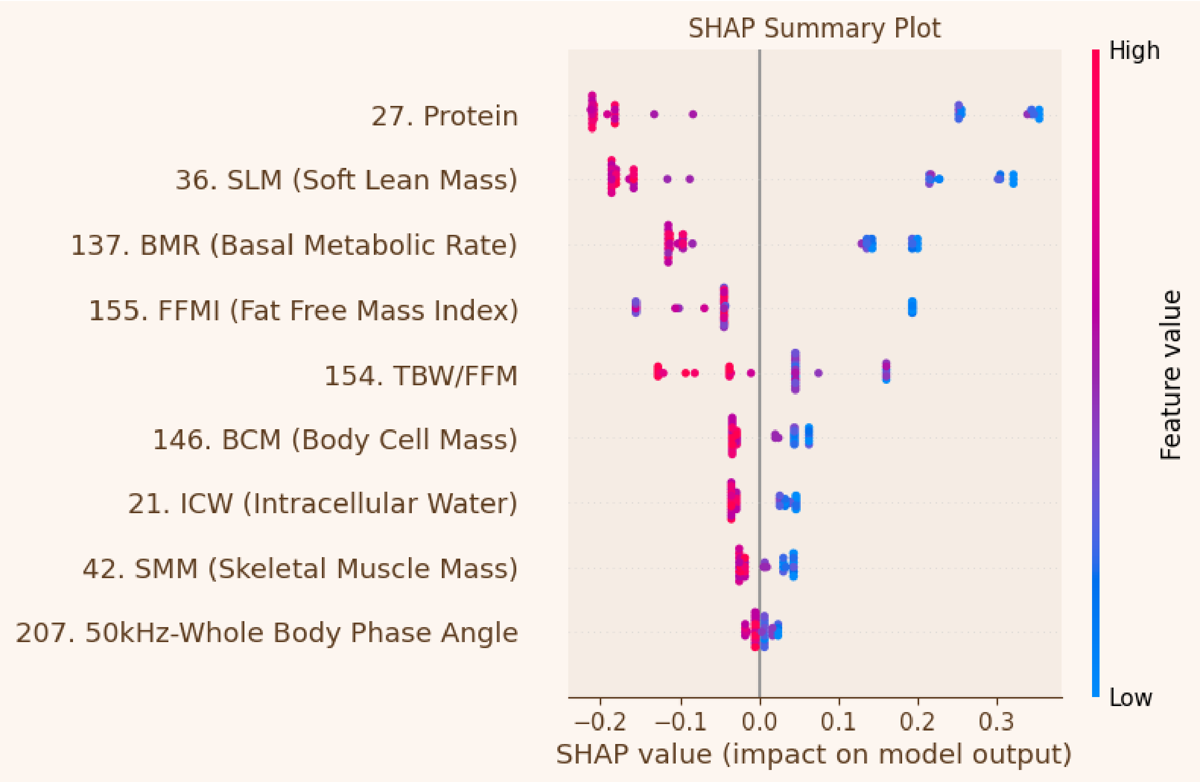


Figure 5.18: SHAP summary plot for the selected body composition features.

mean. Narrow bands suggest the mean is estimated precisely (typically, more participants and less variability); wider bands indicate more uncertainty. Overlapping bands between clusters imply that their mean trajectories are not clearly different at that moment, whereas a clear gap with little or no overlap suggests a more credible difference. Note that these bands describe uncertainty around the *mean*, not the spread of individual children's values.

Figure 5.19 displays the *mean trajectory of BMI percentile* across the three assessment moments (T1–T3). The clusters defined on body composition features retain their relative ordering over time, with limited convergence.

- **Cluster 1** starts at a higher BMI percentile (around 75 at T1), shows a small decline by T2, and a small increase at T3. Confidence bands are comparatively narrow, indicating limited within cluster variability.
- **Cluster 2** remains in the mid-range throughout (mid–60s), with only minor fluctuations across waves. Uncertainty bands are moderate and largely stable.
- **Cluster 3** begins lower (around the 30th percentile), rises at T2 (mid–40s), and partially reverts by T3 (low–30s). This group shows the widest uncertainty, reflecting greater heterogeneity.

Overall, trajectories suggest short-term persistence of baseline profiles: clusters remain separated on average, with Cluster 3 exhibiting the greatest temporal variability.

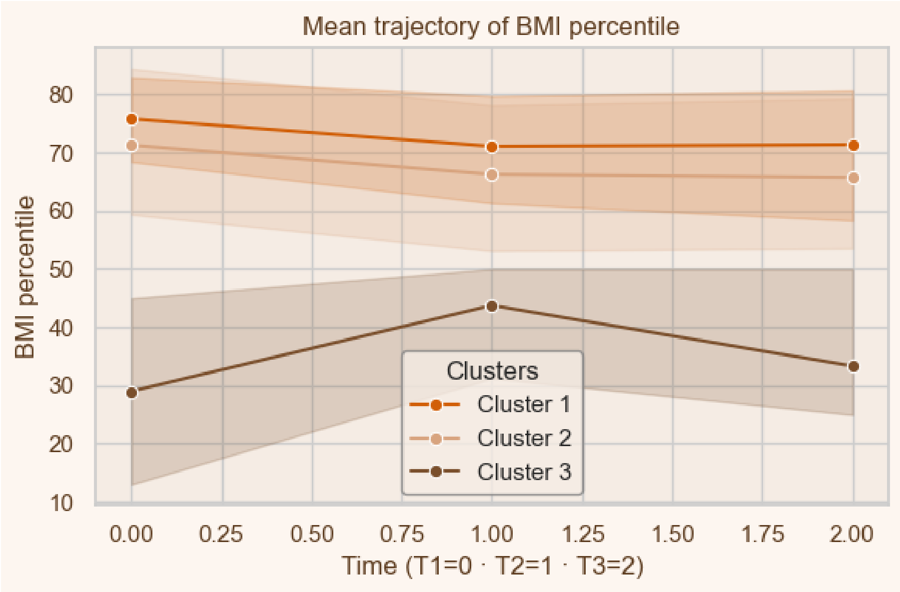


Figure 5.19: Mean trajectory of BMI percentile by cluster across time. Time is coded as T1 $= 0$, T2 $= 1$, T3 $= 2$.

5.7 Summary

This section provides a concise synthesis of the main findings derived from the clustering and subsequent analyses. It integrates results across body composition, behavioural, and contextual domains to highlight how the identified profiles align with children's functional performance, family environment, and perceptions related to health and well-being.

- **Three profiles.** The dendrogram cut results in clusters that are distinguishable in both projection space and distributional summaries, with a moderate Silhouette score (0.34).
- **Drivers of separation.** SHAP and visual profiling converge on lean-mass and cellular–water markers (Protein, SLM, SMM, FFMI, BCM, ICW, TBW/FFM, Phase Angle) as the principal differentiators: lower values characterise the higher-risk group.
- **Cross-domain alignment.** Joint density plots indicate that higher BMI percentiles match with less favourable ranges in cardiorespiratory and functional indices, particularly within the higher-risk cluster (Cluster 3); more favourable ranges align with the lower-risk cluster (Cluster 1).
- **Family context.** Parental indices show directionally consistent trends (higher activity and diet quality, lower screen and sedentary time, and food insecurity) in families of children within the lower-risk cluster.
- **Perceptions.** Drawing-derived indices tend to be more healthful in the lower-risk cluster and less so in the higher-risk cluster, providing a perceptual complement

to objective measures.

- **Temporal stability.** Cluster differences in BMI percentile are broadly maintained across T1–T3, with limited convergence over the observed window.

Overall, the results suggest that clusters defined by body composition map meaningfully onto functional capacity, family context, and children's health-related perceptions, and that these profiles display short-term stability. This integrated picture supports the utility of unsupervised profiles for early obesity risk assessment and for guiding targeted interventions.

6 Discussion

This project involved the development and interpretation of Artificial Intelligence models to estimate obesity risk in school-aged children using multi-source data from the GicO study (body composition, cardiorespiratory/functional fitness, parental questionnaires, and children's drawings). Unsupervised analysis using hierarchical clustering suggested three consistent groups with differences in lean and cell-mass metrics, fitness and lifestyle indices, which broadly persisted across phases T1–T3. Model explanations (SHAP) were useful in highlighted body composition features (e.g., *Protein, Soft Lean Mass*), cardiorespiratory capacity (VO_2), and parental indices (e.g., PAI, DQI, STI) as the most influential markers.

6.1 Interpretation across data domains

6.1.1 Body composition

Body composition provided the strongest signal. In particular, [*Protein, Soft Lean Mass, FFMI, phase angle, ECW/TBW*] featured prominently among top SHAP contributors. Replacing raw BMI with BMI percentiles (derived from Portuguese growth curves) aligned the models with children-oriented standards, avoided adult BMI categories, and reduced age-related confounding in this narrow 6–8-year band. This choice was reflected in improved clusters definition.

6.1.2 Parental questionnaire indices

Parental indices captured behaviours and environments: dietary quality (DQI), sedentary tendencies (STI), physical activity (PAI), sleep, and food access. Notably, higher STI and lower PAI were associated with the higher-risk cluster.

6.1.3 Cross-domain correlation

Combining body composition, fitness, and questionnaire indices in the same plots allowed us to evaluate the data jointly and identify characteristics that help explain the clusters. This cross-domain view revealed patterns not visible within single domains and supported a clearer description of each cluster.

6.2 Clustering and longitudinal trajectories

The three clusters, normal – favourable composition (Cluster 1), normal – average composition (Cluster 2), and higher-risk (Cluster 3), were recovered with a modest but acceptable silhouette (≈ 0.34). Longitudinally, mean BMI percentile trajectories were largely stable over T1–T3, with some children in the higher-risk cluster exhibiting small increases, and the lower-risk group remaining comparatively flat. With few participants and attrition across T1–T3, most children stayed in their starting profile, reinforcing the need for early, family-centred action in line with GicO.

6.3 Practical implications for screening and intervention

In settings similar to the GicO study (e.g schools), the models can support soft triage: flagging children for a personalised conversation rather than a clinical label. Decision thresholds should prioritise sensitivity to minimise missed cases, followed by confirmatory review by the team.

6.4 Strengths

Based on the overall work, we can highlight the following strengths of the analysis:

- Multi-source and family context data: anthropometry, bioimpedance, fitness, and parental environment/behaviour indices, reflecting the real drivers addressed by the GicO study.
- BMI percentiles are an appropriate adiposity metric for children and improved construct validity over adult BMI categories in this age band.
- Hierarchical clustering and SHAP-based interpretability reduce the risk of black-box modelling.

6.5 Limitations

Regarding the limitations, we can say:

- The small cohort (n=34) constrains statistical power, particularly for detecting other patterns.
- Participants are children enrolled in a gymnastics programme in Coimbra-Portugal; findings may not generalise to other sports, regions, or socioeconomic contexts without adaptation.
- Bioimpedance in children has measurement constraints and has assumptions about

hydration; fitness tests can be affected by motivation and learning effects; questionnaire data are self-reported.

- Unmeasured factors (e.g., maturity status, family routines, food marketing exposure) may influence risk.

7 Conclusion

This project was guided by the main objective of: developing AI models to identify determinants of childhood obesity risk by integrating the multi-source GicO datasets and translating evidence into profiles and indicators for early screening in community settings. This objective was successfully met. In particular, we used interpretable AI and targeted feature engineering to explore how parental lifestyles relate to children's body composition, producing actionable profiles and indicators that support prevention in community contexts.

The unsupervised analysis yielded three children's profiles driven primarily by lean-tissue and cellular-water markers. **Cluster 1 (Normal - favourable composition):** reflects a healthier composition with the highest lean-mass and cellular-water, higher phase angle, adiposity generally more favourable; consistent with better metabolic, cardiorespiratory, and functional capacity. **Cluster 2 (Normal - average composition):** intermediate central values with wider dispersion and BMI percentile spanning a broad range, reflecting a normal/transitional composition. **Cluster 3 (Higher-risk):** reflects a less healthy composition with the lowest lean and cellular-water indices, lower phase angle, and higher relative adiposity, aligning with a less favourable metabolic/functional outlook. Collectively, these clusters map a clear progression from optimal to less healthy profiles and suggest targets for action: increase fat-free mass, improve hydration balance, and reduce excess adiposity. Within the GicO framework, combining groups' profiling can help target interventions where they are most needed.

By constructing composite indices based on the parental questionnaire and using unsupervised models, clear associations were identified between:

- parental indices such as physical activity (PAI) and diet quality (DQI), which are associated with healthier body-composition profiles in children;
- parental indices of screen/sedentary time (STI) and food insecurity (FIS), which are associated with a higher risk of obesity.

These findings reinforce the importance of promoting healthy lifestyles within the family as an early intervention tool against childhood obesity.

By converting subjective visual data (children's drawings) into structured indices, it was possible to integrate qualitative perceptions into the analytical model. These indices serve as innovative markers, enriching the quantitative analysis with a psychosocial and imaginative dimension, and offering insights into how children conceptualise health and well-being.

The composite indices convert a 25-minute parental interview into a concise behavioural fingerprint suitable for understanding family context.

This project demonstrates that multi-source data analysed with interpretable AI can add practical value to community programmes tackling childhood obesity risk. Body composition remains a strong signal, but fitness, behaviours, and family context provide complementary information that improves the identification of groups. Incorporating BMI percentiles based on age and sex was essential for paediatric relevance.

7.1 Future work

While the present work provides a comprehensive analysis of multi-source and multi-dimensional data relating to childhood obesity, several directions for future research and data collection remain open. Specifically:

- Future work could extend the current analysis to a broader sample, enabling the development of supervised classification models that predict risk profiles from the integrated dataset. These models could enhance the interpretability and generalisability of the findings obtained from the current unsupervised approach. Additionally, this could be complemented by a visual dashboard integrating the key indicators, allowing each child to be positioned within higher-risk, normal (average composition), or normal (favourable composition) categories according to their body composition, functional fitness, and family context.
- An interesting direction involves combining the cardiorespiratory and functional data with the parental questionnaires to derive new clusters related to lifestyle. The goal would be to group families according to their health behaviours and awareness of healthy lifestyles. Once such clusters are formed, subsequent analyses could assess whether gymnasts from families with healthier lifestyles show better baseline composition and fitness indicators, as well as more pronounced improvements over time.
- Another relevant line of work is to identify which variables are most interrelated. Understanding these correlations could help reduce the dimensionality of the datasets, highlighting a small subset of core indicators that can be easily monitored in future cohorts.
- Although the parental questionnaires include behavioural indicators, detailed quantitative data on children's nutritional intake (e.g., food frequency, caloric intake, or macro-/micronutrient profiles) were not collected. Including such measures in future studies could significantly enhance dietary profiling and its association with obesity indicators.

- Future work could deepen the longitudinal analysis using clustering over time to group children by common change patterns. Apply methods to capture trajectories and predict risk evolution. Include explanations to show what drives changes within a child.

References

[1] "DE-RISKING METABOLIC, ENVIRONMENTAL AND BEHAVIORAL DETERMINANTS OF OBESITY IN CHILDREN, ADOLESCENTS AND YOUNG ADULTS," 2023, pAS GRAS project, Horizon Europe.

[2] P. G. P. Team, "Gico study – annex 1: Research plan and respective stages," Unpublished manuscript, 2024, internal document of the PAS GRAS project.

[3] "Combater a obesidade e o excesso de peso desde criança." [Online]. Available: https://noticias.uc.pt/artigos/combater-a-obesidade-e-o-excesso-de-peso-desde-crianca/

[4] Faculty of Sport Sciences and Physical Education, University of Coimbra, "Faculty of sport sciences and physical education – university of coimbra," https://www.uc.pt/en/fcdef/, 2025, accessed: 2025-07-09.

[5] "Cnc center for neuroscience and cell biology | cnc uc." [Online]. Available: https://cnc.uc.pt/en

[6] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.

[7] InBody Portugal, "Inbody 770 – analisador de composição corporal," https://www.inbody.pt/inbody-770.html, 2025, accessed: 2025-07-09.

[8] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[9] H. Zhao, W. Ding, F. Ye, W. Yuan, and H. Chen, "The clustering analysis on public health data with missing values based on dimension reduction methods," *Proceedings of the 2023 International Conference on Health Big Data and Intelligent Healthcare (ICHIH 2023)*, pp. 45–51, 2023.

[10] F. Sharifi, E. Mohammed, T. Crump, and B. H. Far, "A cluster-based machine learning model for large healthcare data analysis," vol. 1054, pp. 92–106, 2019.

[11] M. Paoletti *et al.*, "Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of chronic obstructive pulmonary disease (copd) phenotypes," *Journal of Biomedical Informatics*, vol. 42, no. 6, pp. 1013–1021, 2009.

[12] F. Sharifi, E. Mohammed, T. Crump, and B. H. Far, "A cluster-based machine learning model for large healthcare data analysis," in *Big Data Innovations and Applications*, M. Younas, I. Awan, and S. Benbernou, Eds. Cham: Springer International Publishing, 2019, pp. 92–106.

[13] U. Niemann *et al.*, "Phenotyping chronic tinnitus patients using self-report questionnaire data: cluster analysis and visual comparison," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.

[14] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[16] U.S. Food and Drug Administration, "Patient-reported outcome measures: Use in medical product development to support labeling claims," Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), U.S. Food and Drug Administration, Silver Spring, MD, Guidance for Industry, Dec. 2009. [Online]. Available: https://www.fda.gov/media/77832/download

[17] J. Zhou, B. Muluneh, Z. Wang, H. Yao, and J. H. Hughes, "Leveraging longitudinal patient-reported outcome trajectories to predict survival in non-small cell lung cancer," *Clinical Cancer Research*, vol. 31, no. 13, pp. 2685–2694, 2025.

[18] W. J. Hou, J. X. Liu, and X. Y. Yan, "A questionnaire data clustering method based on optimized k-modes algorithm," in *Lecture Notes in Computer Science*, vol. 12797, 2021, pp. 326–342.

[19] B. C. Boehmke and B. M. Greenwell, *Hands-On Machine Learning with R*, ser. Chapman & Hall/CRC The R Series. Boca Raton, FL: CRC Press, 2019.

[20] M. Miyahara, "A pilot exploratory study to form subgroups using cluster analysis of family needs survey scores for providing tailored support to parents caring for a population-based sample of 5-year-old children with developmental concerns," *International Journal of Environmental Research and Public Health*, vol. 19, no. 2, p. 744, 2022.

[21] M. Nardo, M. Saisana, A. Saltelli, S. Tarantola, A. Hoffman, and E. Giovannini, *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, 2008. [Online]. Available: https://www.oecd.org/en/publications/handbook-on-constructing-composite-indicators-methodology-and-user-guide_9789264043466-en.html

[22] B. Wang, C. Nong, J. Zhang, L. Deng, W. Li, X. Zhang, and P. Liu, "Prevalence and

associated body composition factors of sarcopenia in community-dwelling older adults," *European journal of medical research*, vol. 29, p. 598, 12 2024. [Online]. Available: https://link.springer.com/articles/10.1186/s40001-024-02185-9https://link.springer.com/article/10.1186/s40001-024-02185-9

[23] Direção-Geral da Saúde, "Circular Normativa Nº 05/DSMIA — Consultas de Vigilância de Saúde Infantil e Juvenil: Atualização das Curvas de Crescimento," https://www.dcc.fc.up.pt/~ines/aulas/1516/DM1/UCMF-explanations-and-tables/imc_sandra.pdf, February 2006, contact: Divisão de Saúde Materna, Infantil e dos Adolescentes.

[24] D. J. e. a. Buysse, "The pittsburgh sleep quality index: a new instrument for psychiatric practice and research," *Psychiatry Research*, vol. 28, no. 2, p. 193–213, 1991.

[25] B. E. Ainsworth, W. L. Haskell, S. D. Herrmann, and et al., "2011 compendium of physical activities: A second update of codes and met values," *Medicine & Science in Sports & Exercise*, vol. 43, no. 8, pp. 1575–1581, 2011.

[26] E. Stamatakis, J. Gale, A. Bauman, U. Ekelund, M. Hamer, and D. Ding, "Sitting time, physical activity, and risk of mortality in adults," *Journal of the American College of Cardiology*, vol. 73, no. 16, pp. 2062–2072, 2019.

[27] P. M. Guenther, S. I. Kirkpatrick, J. Reedy, and et al., "The healthy eating index–2010 is a valid and reliable measure of diet quality," *Journal of Nutrition*, vol. 144, no. 3, pp. 399–407, 2013.

[28] A. Trichopoulou, T. Costacou, C. Bamia, and D. Trichopoulos, "Adherence to a mediterranean diet and survival in a greek population," *New England Journal of Medicine*, vol. 348, no. 26, pp. 2599–2608, 2003.

[29] T. Ballard, A. Kepple, and C. Cafiero, "The food insecurity experience scale: Development of a global standard for monitoring hunger worldwide," FAO Statistics Division, Rome, Tech. Rep., 2013. [Online]. Available: http://www.fao.org/economic/ess/ess-fs/voices/en/

[30] C.-C. Teng and C. Chih, "Sustainable food literacy: A measure to promote sustainable diet practices," *Sustainable Production and Consumption*, vol. 30, pp. 776–786, 2022.

[31] W. H. Organization, "Urban green space interventions and health: A review of impacts and effectiveness," https://www.euro.who.int/__data/assets/pdf_file/0010/337690/Factsheet-Green-spaces-and-health-ENG.pdf, 2017, accessed 1 July 2025.

[32] R. Santos, C. Silva, B. Rodrigues, and J. Tavares, "Adaptação cultural e validação da versão portuguesa do pittsburgh sleep quality index (psqi)," *Revista de Psiquiatria Clínica*, vol. 41, no. 4, pp. 99–105, 2014.

[33] P. F. Lovibond and S. H. Lovibond, *Manual for the Depression Anxiety Stress Scales*, 2nd ed. Sydney: Psychology Foundation of Australia, 1995.

[34] P. Braveman, C. Cubbin, S. Egerter, D. R. Williams, and E. Pamuk, “Socioeconomic disparities in health in the united states: What the patterns tell us,” *American Journal of Public Health*, vol. 100, no. S1, pp. S186–S196, 2010.

[35] M. Barbosa-Silva, A. Barros, J. Wang, S. Heymsfield, and R. Pierson, “Bioelectrical impedance analysis: population reference values for phase angle by age and sex,” *The American journal of clinical nutrition*, vol. 82, no. 1, pp. 49–52, 2005.

[36] GeeksforGeeks, “Hierarchical clustering in machine learning,” https://www.geeksforgeeks.org/machine-learning/hierarchical-clustering/, 2023, accessed: 2025-09-06.

[37] S. Developers, “Scipy cluster hierarchy documentation,” https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html, 2024, accessed: 2025-09-06.

[38] B. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed. Chichester, UK: Wiley, 2011.

Appendices

Appendix A - Project Proposal

The following appendix includes the project proposal.

Instituto Superior de Engenharia
Politécnico de Coimbra

Mestrado em Engenharia Informática
Proposta de Projeto

Ano Letivo de 2024/2025

Título	Modelos Inteligentes para Análise e Predição de Riscos de Obesidade
Aluno	Sandra Nathaly Aguilar Alvarez
Orientadores	Francisco Pereira (DEIS-ISEC) António Malta Lopes da Cruz, Ph.D., Centro de Neurociências e Biologia Celular UC, antoniocruz@cnc.uc.pt

SUMÁRIO

Neste projeto pretende-se desenvolver e validar modelos de análise inteligente de dados, para estimar o risco de obesidade e suas complicações a partir de dados biológicos, dados bioquímicos e respostas a questionários. Os métodos serão desenvolvidos a partir de datasets disponibilizados pelos parceiros do projeto PAS GRAS. Os modelos deverão suportar a identificação e análise de padrões relevantes para obter conhecimento sobre as causas subjacentes da obesidade e prever possíveis complicações futuras.

1. ÂMBITO

O projeto "PAS GRAS: redução de riscos metabólicos, determinantes ambientais e comportamentais da obesidade em crianças, adolescentes e jovens adultos" é financiado pelo Horizonte Europa e envolve 8 países. Uma das componentes do projeto incide no desenvolvimento de métodos para análise e avaliação de riscos de obesidade através dos conjuntos de dados recolhidos pelos vários grupos envolvidos. A equipa é interdisciplinar com atuação nas áreas das ciências da vida, saúde, desporto, educação e comunicação.

2. OBJETIVOS

O presente projeto pretende atingir os seguintes objetivos genéricos:

- Integração na equipa de investigação do projeto PAS GRAS
- Pré-processamento e análise exploratória dos dados recolhidos pelas equipas de investigação
- Criação e validação de modelos não supervisionados que identifiquem padrões relevantes sobre as causas subjacentes à obesidade
- Criação e validação de modelos supervisionados que sirvam como preditores de possíveis complicações futuras

Instituto Superior de Engenharia
Politécnico de Coimbra

3. PLANO DE TRABALHOS

O projeto consistirá nas seguintes atividades e respetivas tarefas:

T1 *– Estado da arte –* Revisão da bibliografia e estudo do domínio do problema
T2 *– Análise Exploratória –* Compreensão, pré-processamento e análise exploratória de dados
T3 *– Análise não supervisionada –* Criação e validação de modelos não supervisionados
T4 *– Perfilagem de grupos–* interpretação dos clusters descobertos, construção de perfis de risco, alinhamento com conhecimento especializado
T5 - *Documentação* – Produção de documentação

Serão produzidos os seguintes documentos ou materiais:

E1 *– Modelos não supervisionados –* Modelos inteligentes de análise não supervisionada.
E2 *– Modelos supervisionados –* Protótipos de modelos inteligentes de análise supervisionada
E3 *– Resumo para Workshop Intercalar –* Resumo e Apresentação PPT para workshop intercalar
E4 *– Relatório final do projeto*
E5 *– Colaboração com a equipa de investigação do PAS GRAS na produção de documentação científica*

A calendarização prevista para os trabalhos e entregas será a seguinte:

	Set		Out		Nov		Dez		Jan		Fev		Mar		Abr		Mai		Jun		Jul	
Tarefas																						
T1																						
T2																						
T3																						
T4																						
T5																						
Entregas																						
E1																X						
E2																				X		
E3														X								
E4																						X

4. METODOLOGIA DE TRABALHO

O trabalho será realizado preferencialmente no DEIS-ISEC. Periodicamente, será necessário participar em reuniões e outras atividades do projeto PAS GRAS nas instalações do CNC-UC.

Appendix B - Parental Questionnaire (Original Format)

The following appendix includes the full PDF of the original questionnaire administered to parents.

QUESTIONÁRIO ONLINE (a cada um dos pais)

Um futuro mais saudável para todos

Combater obesidade com ginástica

É convidado(a) a participar neste estudo que visa avaliar a eficácia de programas que combinam exercício, alimentação, e conteúdos de ciência e saúde, na prevenção e tratamento de excesso de peso e obesidade em crianças em idade escolar.

Este estudo está integrado no projeto europeu PAS GRAS e junta uma equipa interdisciplinar única, que beneficia do envolvimento das crianças e respetivas famílias, investigadores, profissionais de saúde, entre outros, com o objetivo de reduzir a curva da obesidade, especialmente nas populações mais jovens, e proporcionar programas para uma saúde e bem-estar global e ao longo da vida.

A sua participação envolve responder a um conjunto de perguntas associadas ao seu agregado familiar. As suas respostas serão confidenciais, seguindo as diretivas nacionais e europeias relativas à proteção de dados, e os resultados serão analisados apenas para investigação. O tempo de preenchimento do questionário será inferior a 20-25 minutos. Se desejar obter informação adicional, poderá contactar-nos para o e-mail: pas.gras@uc.pt

Antes de começar, indique o código do estudo que lhe foi atribuído, **ID** ________________________

1. DESPORTO E ATIVIDADE FÍSICA

1.1. Pratica algum **desporto** com regularidade? Não ☐ Sim ☐

1.1.1. Se respondeu **sim**, quantas vezes por semana? ____ vezes. E quantas horas por sessão/treino? ____ horas

1.1.2. Se respondeu **sim**, diga qual(is) o(s) desporto(s): ____

1.2. Relativamente a **atividade física**, indique (X) o tempo em minutos (min) que dedica a:

	Nunca	Menos de 30 min por semana	30 a 60 min por semana	60 a 90 min por semana	Mais de 90 min por semana
Caminhar	☐	☐	☐	☐	☐
Andar de bicicleta	☐	☐	☐	☐	☐
Cuidar da casa	☐	☐	☐	☐	☐
Atividades no campo	☐	☐	☐	☐	☐
Atividades horta ou jardim	☐	☐	☐	☐	☐
Dançar	☐	☐	☐	☐	☐
Outra	☐	☐	☐	☐	☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

2. AMBIENTE DIGITAL EM CASA

2.1. Que equipamentos digitais tem em **casa**? E quais usa habitualmente no **quarto**?

Em casa		No quarto	
Televisão	☐	Televisão	☐
Computador (fixo e/ou portátil)	☐	Computador (fixo e/ou portátil)	☐
Tablet	☐	Tablet	☐
Telemóvel	☐	Telemóvel	☐
Consola de jogos (GameBoy, Nintendo)	☐	Consola de jogos (GameBoy, Nintendo)	☐
Blu-ray/DVD/CD/Cassetes de vídeo	☐	Blu-ray/DVD/CD/Cassetes de vídeo	☐
Outros	☐	Outros	☐
Se outros, quais?		Se outros, quais?	

2.2. Tem internet em casa? Não ☐ Sim ☐

Se **sim**, quem utiliza habitualmente a internet? ______________________________

2.3. Relativamente a equipamentos digitais, indique com que idade é que **utilizou pela primeira vez**:

2.3.1. **Televisão:** ____ ano(s) OU nunca assistiu ☐
2.3.2. **Computador:** ____ ano(s) OU nunca usou ☐
2.3.3. **Consola de videojogo**: ____ ano(s) OU nunca usou ☐
2.3.4. **Tablet**: ____ ano(s) OU nunca usou ☐
2.3.5. **Telemóvel**: ____ ano(s) OU nunca usou ☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

2.4. Indique (X) o **tempo de utilização** dos seguintes aparelhos electrónicos:

COMPUTADOR (fixo e/ou portátil)					
Num dia de semana (entre 2ª e 6ª feira)					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao sábado					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao domingo					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐

TABLET					
Num dia de semana (entre 2ª e 6ª feira)					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao sábado					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao domingo					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐

TELEMÓVEL					
Num dia de semana (entre 2ª e 6ª feira)					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao Sábado					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao Domingo					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐

TELEVISÃO					
Num dia de semana (entre 2ª e 6ª feira)					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao Sábado					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao Domingo					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

CONSOLA DE VÍDEO-JOGOS					
Num dia de semana (entre 2ª e 6ª feira)					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao Sábado					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐
Ao Domingo					
Nenhuma ☐	Até 1 hora ☐	1 a 2 horas ☐	2 a 3 horas ☐	3 a 4 horas ☐	Mais de 4 horas ☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

3. HÁBITOS ALIMENTARES

3.1. Quantas refeições faz por dia? ______

3.1.1.Indique (X) qual a **frequência** das seguintes refeições:

	Nunca ou raramente	1-3 vezes por mês	1 dia por semana	2-4 dias por semana	5-6 dias por semana	Todos os dias
Pequeno-almoço	☐	☐	☐	☐	☐	☐
Lanche da manhã	☐	☐	☐	☐	☐	☐
Almoço	☐	☐	☐	☐	☐	☐
Lanche da tarde	☐	☐	☐	☐	☐	☐
Jantar	☐	☐	☐	☐	☐	☐
Ceia	☐	☐	☐	☐	☐	☐

3.2. Indique (X) o **local** e a **frequência** das refeições principais (pequeno-almoço, almoço ou jantar):

	Nunca ou raramente	1-3 vezes por mês	1 dia por semana	2-4 dias por semana	5-6 dias por semana	Todos os dias
Cozinhadas/preparadas em casa	☐	☐	☐	☐	☐	☐
Restaurantes (p.e., tradicionais portugueses, brasileiros, asiáticos, africanos...)	☐	☐	☐	☐	☐	☐
Restaurantes *fast-food* (p.e., Taco Bell McDonald's, Pizza Hut, Burger King, KFC...)	☐	☐	☐	☐	☐	☐
Cantinas/refeitórios (p.e., escolas)	☐	☐	☐	☐	☐	☐
Pastelarias/cafetarias	☐	☐	☐	☐	☐	☐
Pedidos por aplicação, *take away*, no local (p.e., Uber, Bolt Glovo)	☐	☐	☐	☐	☐	☐

3.2.1. Tem por hábito cozinhar? __ Com que frequência? ____

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

3.3. Indique (X) qual a **frequência** com que consome os seguintes tipos de alimentos?

	Nunca ou raramente	1-3 vezes por mês	1 dia por semana	2-4 dias por semana	5-6 dias por semana	1 vez por dia	Mais do que 1 vez por dia
Bolos/pastéis (p.e., Donuts, *Bollycao*)	☐	☐	☐	☐	☐	☐	☐
Cereais de pequeno-almoço (p.e., Corn Flakes, Golden Grahams, Nesquik, Kellog's, Alpen)	☐	☐	☐	☐	☐	☐	☐
Bolachas e biscoitos	☐	☐	☐	☐	☐	☐	☐
Chocolates, gomas e rebuçados	☐	☐	☐	☐	☐	☐	☐
Batatas-fritas (de pacote), *snacks*, salgados, pipocas	☐	☐	☐	☐	☐	☐	☐
Refrigerantes ou bebidas açucaradas, sumos de pacote/garrafa (p.e., Fanta, Coca-cola, Compal)	☐	☐	☐	☐	☐	☐	☐
Bebidas alcoólicas (como cerveja, vinho)	☐	☐	☐	☐	☐	☐	☐
Fast-food (como pizza, hambúrgueres, *nuggets* de frango)	☐	☐	☐	☐	☐	☐	☐
Processados (como fiambre, salsichas, bacon, presunto)	☐	☐	☐	☐	☐	☐	☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329
Funded by the European Union

3.4. Indique (X) qual a **frequência** com que consome os seguintes tipos de alimentos?

	Nunca ou raramente	1-3 vezes por mês	1 dia por semana	2-4 dias por semana	5-6 dias por semana	1 vez por dia	Mais do que 1 vez por dia
Pão, cereais e tubérculos (como batata), arroz ou massa	☐	☐	☐	☐	☐	☐	☐
Lacticínios (leite branco, queijo, iogurtes)	☐	☐	☐	☐	☐	☐	☐
Gordura vegetal (óleo, margarina)	☐	☐	☐	☐	☐	☐	☐
Gordura vegetal (azeite)	☐	☐	☐	☐	☐	☐	☐
Gordura animal (manteiga, banha)	☐	☐	☐	☐	☐	☐	☐
Carnes vermelhas (vaca, porco, coelho)	☐	☐	☐	☐	☐	☐	☐
Carnes de aves (como frango, peru...)	☐	☐	☐	☐	☐	☐	☐
Peixe	☐	☐	☐	☐	☐	☐	☐
Ovos	☐	☐	☐	☐	☐	☐	☐
Frutos oleaginosos (nozes, amêndoas, avelãs, amendoins)	☐	☐	☐	☐	☐	☐	☐
Leguminosas (ervilhas, favas, lentilhas, feijão)	☐	☐	☐	☐	☐	☐	☐
Fruta (como laranja, maçã, pera, banana)	☐	☐	☐	☐	☐	☐	☐
Cogumelos silvestres	☐	☐	☐	☐	☐	☐	☐
Cogumelos cultivados	☐	☐	☐	☐	☐	☐	☐
Legumes crus / saladas	☐	☐	☐	☐	☐	☐	☐
Legumes cozidos	☐	☐	☐	☐	☐	☐	☐
Sopa	☐	☐	☐	☐	☐	☐	☐

3.5. Tem por hábito consumir alimentos da época (sazonais)? Não ☐ Sim ☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

3.5.1. Se **sim**, quais os **mais** consumido(s): ____

4. ACESSO A ALIMENTOS

4.1. Indique (X) qual a **frequência com que compra** a sua comida:

	Nunca ou raramente	1-3 vezes por mês	1 vez por semana	2-3 vezes por semana	4-6 vezes por semana	Todos os dias
Hipermercado	☐	☐	☐	☐	☐	☐
Supermercado, minimercado, mercearia	☐	☐	☐	☐	☐	☐
Mercado	☐	☐	☐	☐	☐	☐
Outro	☐	☐	☐	☐	☐	☐

4.1.1. (Com base nos resultados da pergunta anterior) Por que prefere comprar comida no/a (local que indicou) ________?

4.2. Indique (X) a importância ao **escolher o local onde compra** alimentos (sendo 1 o menos importante e 6 o mais importante):

	1	2	3	4	5	6
Distância	☐	☐	☐	☐	☐	☐
Preço	☐	☐	☐	☐	☐	☐
Qualidade dos produtos	☐	☐	☐	☐	☐	☐
Outro(s)	☐	☐	☐	☐	☐	☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

4.3. Indique (X) a importância para **escolher os alimentos que compra** (sendo 1 o menos importante e 6 o mais importante):

	1	2	3	4	5	6
Qualidade	☐	☐	☐	☐	☐	☐
Facilidade de preparação	☐	☐	☐	☐	☐	☐
Frescura (não ultra-processado)	☐	☐	☐	☐	☐	☐
Outro(s)	☐	☐	☐	☐	☐	☐

4.4. Sobre o aceso a alimentos, para cada pergunta, responda (X), o que melhor se adequa à sua situação. Não há respostas certas ou erradas. Não leve muito tempo a indicar a resposta em cada afirmação.

Nos últimos 6 meses…	**Sim**	**Não**	**Não sabe**	**Não responde**
1. Sentiu preocupação com a comida acabar em casa, antes de ter dinheiro para comprar mais?	☐	☐	☐	☐
2. A comida em sua casa acabou antes de ter dinheiro para comprar mais?	☐	☐	☐	☐
3. A sua família ficou sem dinheiro suficiente para conseguir ter uma alimentação saudável e variada?	☐	☐	☐	☐
4. A sua família teve de consumir apenas os alimentos que ainda tinha em casa por ter ficado sem dinheiro para comprar mais?	☐	☐	☐	☐
5. Saltou alguma refeição por não ter dinheiro suficiente para comprar comida?	☐	☐	☐	☐
6. Alguém adulto da sua família comeu menos por não ter dinheiro suficiente para comprar comida?	☐	☐	☐	☐
7. Alguém adulto da sua família sentiu fome mas não comeu por falta de dinheiro para comprar comida?	☐	☐	☐	☐
8. Alguém adulto da família ficou um dia inteiro sem comer, com apenas uma refeição ao longo do dia, por não ter dinheiro suficiente para comprar comida?	☐	☐	☐	☐
As crianças/adolescentes da sua família... (agregados familiares com crianças até aos 18 anos de idade)				
9. Não conseguiram ter uma alimentação saudável e variada (i.e. legumes, peixe, fruta,…) por falta de dinheiro?	☐	☐	☐	☐
10. Tiveram de consumir apenas alguns alimentos que ainda tinham em casa por terem ficado sem dinheiro?	☐	☐	☐	☐
11. Comeram menos do que deviam por não ter dinheiro para comprar comida?	☐	☐	☐	☐
12. Tiveram menos comida às refeições por não ter dinheiro suficiente para comprar mais?	☐	☐	☐	☐
13. Deixaram de fazer alguma refeição por não haver dinheiro suficiente para comprar comida?	☐	☐	☐	☐
14. Sentiram fome, mas não comeram por falta de dinheiro para comprar comida?	☐	☐	☐	☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

5. ALIMENTAÇÃO SUSTENTÁVEL

Indique (X) o que melhor se adequa à sua situação (sendo 1 discordo totalmente e 7 concordo plenamente). Não há respostas certas ou erradas. Não leve muito tempo a indicar a resposta em cada afirmação.

ALIMENTAÇÃO SUSTENTÁVEL	1	2	3	4	5	6	7
L1. Entendo e sei aplicar práticas de higiene e segurança alimentar (p.e., lavar bem as mãos, ter unhas curtas e limpas, uso avental sempre que necessário)	☐	☐	☐	☐	☐	☐	☐
L2. Consigo distinguir entre alimentos saudáveis e não saudáveis	☐	☐	☐	☐	☐	☐	☐
L3. Sei onde encontrar informações sobre alimentação e nutrição	☐	☐	☐	☐	☐	☐	☐
L4. Conheço o valor nutricional dos alimentos	☐	☐	☐	☐	☐	☐	☐
L5. Sei onde adquirir alimentos bons para a saúde humana e para o meio ambiente	☐	☐	☐	☐	☐	☐	☐
L6. Sei de onde vem a comida (entendo e sei o que são cadeias alimentares)	☐	☐	☐	☐	☐	☐	☐
L7. Conheço o impacto do consumo de alimentos para a sustentabilidade ecológica e ambiental (p.e., alimentos da época, alimentos de origem local, mercados locais)	☐	☐	☐	☐	☐	☐	☐
L8. Conheço o impacto da produção de alimentos para a diversidade biológica (p.e., uso sustentável dos recursos naturais, terrestres, marinhos e de água doce)	☐	☐	☐	☐	☐	☐	☐
L9. Conheço o impacto da produção de alimentos para as alterações climáticas (p.e., libertação de gases de efeito estufa, poluição)	☐	☐	☐	☐	☐	☐	☐
L10. Sei o que significa "alimento sustentável" (alimento que é “amigo” da saúde humana, ecologia e meio ambiente)	☐	☐	☐	☐	☐	☐	☐
L11. Sei o que significa "bem-estar animal"	☐	☐	☐	☐	☐	☐	☐
L12. Sei o que significa "uma só saúde" (abordagem colaborativa, com o objetivo de alcançar resultados de saúde ideais, reconhecendo a interconexão entre pessoas, animais, plantas, fungos e o seu ambiente partilhado)	☐	☐	☐	☐	☐	☐	☐
COMPETÊNCIAS ALIMENTARES	**1**	**2**	**3**	**4**	**5**	**6**	**7**
CA1. Penso sobre o que comprar antes de comprar	☐	☐	☐	☐	☐	☐	☐
CA2. Sou capaz de armazenar alimentos de forma adequada e segura	☐	☐	☐	☐	☐	☐	☐
CA3. Sou capaz de utilizar utensílios e eletrodomésticos comuns de cozinha	☐	☐	☐	☐	☐	☐	☐
CA4. Sou capaz de preparar e cozinhar alimentos com ingredientes básicos/disponíveis	☐	☐	☐	☐	☐	☐	☐
CA5. Sou capaz de adaptar receitas com base nos alimentos disponíveis	☐	☐	☐	☐	☐	☐	☐
CA6. Sou capaz de improvisar refeições com base nos alimentos disponíveis (incluindo sobras)	☐	☐	☐	☐	☐	☐	☐
INTENÇÃO DE AÇÃO E ESTRATÉGIAS DE AÇÃO	**1**	**2**	**3**	**4**	**5**	**6**	**7**
IA1. Pretendo praticar uma alimentação saudável e sustentável	☐	☐	☐	☐	☐	☐	☐
IA2. Encorajarei outras pessoas a adotarem dietas saudáveis e sustentáveis	☐	☐	☐	☐	☐	☐	☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

IA3. Vou ajustar e praticar uma alimentação saudável e sustentável	☐	☐	☐	☐	☐	☐	☐
EA1. Faço compras a pensar na minha 'pegada ecológica'	☐	☐	☐	☐	☐	☐	☐
EA2. Encorajo outras pessoas a reduzirem o desperdício alimentar	☐	☐	☐	☐	☐	☐	☐
EA3. Encorajo outras pessoas a terem uma alimentação saudável e sustentável	☐	☐	☐	☐	☐	☐	☐
EA4. Apoio campanhas associadas a alimentação sustentável ou zero desperdício de alimentos	☐	☐	☐	☐	☐	☐	☐
ATITUDES	**1**	**2**	**3**	**4**	**5**	**6**	**7**
AT1. Acredito que posso ter uma alimentação saudável e sustentável	☐	☐	☐	☐	☐	☐	☐
AT2. Acredito que o esforço coletivo dos consumidores contribui para uma alimentação saudável e para a sustentabilidade ecológica e ambiental	☐	☐	☐	☐	☐	☐	☐
AT3. Acredito que o esforço a longo prazo das indústrias contribuirá para melhorar para uma alimentação saudável e para a sustentabilidade ecológica e ambiental	☐	☐	☐	☐	☐	☐	☐
AT4. Acredito que as políticas e legislação incentivam a adopção de uma alimentação saudável e a sustentabilidade ecológica e ambiental	☐	☐	☐	☐	☐	☐	☐

Adaptado de Chih-Ching Teng e Chueh Chih (2022) https://doi.org/10.1016/j.spc.2022.01.008

6. AMBIENTE NA ÁREA DE RESIDÊNCIA

6.1. Indique (X) qual a distância de sua casa ao espaço verde mais próximo (considere parques urbanos, bosques e matas de uso público, parques ribeirinhos, jardins públicos e hortas comunitárias):

Até 300 m	300 m – 2000 m	2 km – 5 km	Mais de 5 km
☐	☐	☐	☐

6.2. Indique (X) como considera o **ambiente na área da sua residência** (área em que se desloca a pé):

	Muito bom	**Bom**	**Razoável**	**Mau**	**Muito mau**	**Não sabe/ Não responde**
Qualidade do ar (a presença ou ausência de fumos e mau cheiro associados a trânsito, indústrias)	☐	☐	☐	☐	☐	☐
Espaços públicos de **lazer e recreio** (como praças, parques para skate, circuito de manutenção)	☐	☐	☐	☐	☐	☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

Espaços verdes de fruição de contacto com a **natureza** (como parques urbanos, bosques e matas de acesso público, parques ribeirinhos, corredores verdes, jardins públicos, hortas comunitárias)	☐	☐	☐	☐	☐	☐
Arborização urbana (considere a presença de árvores nas ruas e espaços públicos)	☐	☐	☐	☐	☐	☐
Limpeza e **manutenção urbana** (considere a recolha de lixo, o estado de equipamentos, como bancos, bebedouros, acesso a WC, jardinagem)	☐	☐	☐	☐	☐	☐
Conforto e segurança a **andar pé** (considere o piso pedonal, a presença de árvores)	☐	☐	☐	☐	☐	☐

6.3. Indique (X) qual a **frequência** com que **frui de contacto com a natureza e espaços verdes**.

	Nunca ou raramente	**1-3 vezes por mês**	**1 dia por semana**	**2-4 dias por semana**	**5-6 dias por semana**	**Todos os dias**
Parque com atividades de recreio (como campo de jogos)	☐	☐	☐	☐	☐	☐
Parque sem atividades de recreio	☐	☐	☐	☐	☐	☐
Espaços florestais, matas de acesso público, corredores verdes	☐	☐	☐	☐	☐	☐
Parques ribeirinhos	☐	☐	☐	☐	☐	☐
Jardins públicos	☐	☐	☐	☐	☐	☐
Hortas comunitárias	☐	☐	☐	☐	☐	☐
Outra. Qual?	☐	☐	☐	☐	☐	☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

7. QUANLIDADE DO SONO*

Não há respostas certas ou erradas. Não leve muito tempo a indicar a resposta em cada afirmação.

7.1. Hábitos ao deitar e duração do sono, durante o mês passado:

Horas a que deita na maioria das vezes		Horas a que se levanta na maioria das vezes	
De 2ª a 6ª feira	Ao fim de semana	De 2ª a 6ª feira	Ao fim de semana

Quanto tempo (em minutos) demorou para adormecer na maioria das vezes		Quantas horas de sono por noite dormiu (pode ser diferente do número de horas que ficou na cama)	
De 2ª a 6ª feira	Ao fim de semana	De 2ª a 6ª feira	Ao fim de semana

7.2. Durante o mês passado, indique (X) quantas vezes teve problemas para dormir por causa de:

	Nunca	Menos de 1 vez por semana	1 ou 2 vezes por semana	3 vezes por semana ou mais
Demorar mais de 30 minutos para adormecer	☐	☐	☐	☐
Acordar ao meio da noite ou de manhã muito cedo	☐	☐	☐	☐
Levantar-se para ir à casa de banho	☐	☐	☐	☐
Ter dificuldade para respirar	☐	☐	☐	☐
Tossir ou ressonar alto	☐	☐	☐	☐
Sentir muito frio	☐	☐	☐	☐
Sentir muito calor	☐	☐	☐	☐
Ter sonhos maus ou pesadelos	☐	☐	☐	☐
Outra razão				
Durante o mês passado, quantas vezes teve problemas para dormir por esta razão	☐	☐	☐	☐

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

7.3. Durante o mês passado, indique (X) como classificaria a qualidade do seu sono:

Muito boa	Boa	Razoável	Má	Muito má	Não sabe/ Não responde
☐	☐	☐	☐	☐	☐

7.4. Durante o mês passado, indique (X):

	Nunca	Menos de 1 vez por semana	1 ou 2 vezes por semana	3 vezes por semana ou mais
Tomou algum medicamento para dormir receitado pelo médico, ou indicado por outra pessoa (amigo, familiar, farmacêutico), ou mesmo por sua iniciativa	☐	☐	☐	☐
Teve problemas em ficar acordado durante as refeições, enquanto conduzia, ou enquanto participava nalguma atividade social	☐	☐	☐	☐
Sentiu pouca vontade ou falta de entusiasmo para realizar as suas atividades diárias	☐	☐	☐	☐

*Qualidade do sono de Pittsburgh (psqi-pt) – versão portuguesa

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

8. COMO SE SENTE*

Indique (X) o quanto cada afirmação se aplicou a si durante o **mês passado**.
Não há respostas certas ou erradas. Não leve muito tempo a indicar a resposta em cada afirmação.

	Não se aplicou nada a mim	**Aplicou-se a mim algumas vezes**	**Aplicou-se a mim muitas vezes**	**Aplicou-se a mim a maior parte das vezes**
1. Tive dificuldades em acalmar-me	☐	☐	☐	☐
2. Senti a minha boca seca	☐	☐	☐	☐
3. Não consegui sentir nenhum sentimento bom	☐	☐	☐	☐
4. Senti dificuldades em respirar	☐	☐	☐	☐
5. Tive dificuldade em tomar iniciativa para fazer coisas	☐	☐	☐	☐
6. Tive tendência a reagir em demasia em determinadas situações	☐	☐	☐	☐
7. Senti tremores (por ex., nas mãos)	☐	☐	☐	☐
8. Senti-me nervoso(a)	☐	☐	☐	☐
9. Preocupei-me com situações em que podia entrar em pânico e fazer figura ridícula	☐	☐	☐	☐
10. Senti que não tinha nada a esperar do futuro	☐	☐	☐	☐
11. Dei por mim a ficar agitado(a)	☐	☐	☐	☐
12. Senti dificuldade em relaxar	☐	☐	☐	☐
13. Senti desânimo e tristeza	☐	☐	☐	☐
14. Estive intolerante em relação a qualquer coisa que me impedisse de terminar aquilo que estava a fazer	☐	☐	☐	☐
15. Senti-me quase a entrar em pânico	☐	☐	☐	☐
16. Não fui capaz de ter entusiasmo por nada	☐	☐	☐	☐
17. Senti que não tinha muito valor como pessoa	☐	☐	☐	☐
18. Senti que, por vezes, estava sensível	☐	☐	☐	☐
19. Senti alterações no meu coração sem fazer exercício físico	☐	☐	☐	☐
20. Senti-me assustado(a) sem ter tido uma boa razão para isso	☐	☐	☐	☐
21. Senti que a vida não tinha sentido	☐	☐	☐	☐

*Escala de depressão, ansiedade e stress (eads-21) (http://www2.psy.unsw.edu.au/groups/dass/)

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

9. INFORMAÇÃO SOCIODEMOGRÁFICA

9.1. **Sexo**: Mulher ____ Homem ____ Prefiro não responder ____9.1.1. **Data de nascimento**: _____/____/_____

9.2. As perguntas que se seguem são sobre o **peso e altura**.

9.2.1. Quanto mede? ____m 9.2.2. Quanto pesa? ____kg 9.2.3. Prefiro não responder _____

9.3. As perguntas que se seguem são sobre o **local de residência e pessoas que partilham a casa consigo**:

9.3.1. **Freguesia**: ______________ **Concelho**: ____________ **Distrito**: ___________ **Código postal**: _____-____

9.3.2. Indique (número), quantas pessoas partilham a casa consigo e de que idade:

Até aos 4 anos	Dos 5 aos 17 anos	Dos 18 aos 64 anos	Mais de 65 anos

9.4. **Escolaridade**, indique (X) o nível de ensino que completou:

Sem nível de escolaridade completo ☐
1º ciclo do Ensino Básico (4 anos)/Antiga 4ª classe ☐
2º ciclo do Ensino Básico (6 anos) /Antiga 6ªclasse/Ciclo Preparatório ☐
3º ciclo do Ensino Básico (9 anos)/Curso Geral dos Liceus ☐
Ensino Secundário e pós-secundário (12 anos)/ Curso Complementar dos Liceus ☐
Licenciatura ☐
Mestrado ☐
Doutoramento ☐
Prefere não responder ☐
Outro (indique qual): ☐
____

9.5. Está a trabalhar neste momento?

☐ Sim
☐ Não. Desempregado (desde _____/____/_____)
☐ Aposentado. Qual a última profissão?___________________________________

9.5.1. Se está a trabalhar, qual a **profissão**?
Por favor diga exatamente aquilo que faz, por exemplo, funcionário público, especifique a sua atividade e categoria)

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

9.5.2. Indique (X) qual o grau de intensidade física da sua profissão:

Grande esforço físico	Esforço físico médio	Trabalho muito sedentário
☐	☐	☐

9.6. Qual é o seu meio de transporte habitual nas suas atividades diárias? ________________________

9.7. De uma maneira geral, diria que a sua **saúde** é:

Excelente	Muito boa	Boa	Razoável	Fraca	Muito fraca	Não sabe/ Não responde
☐	☐	☐	☐	☐	☐	☐

9.7.1. Foi-lhe diagnosticado alguma destas condições? Indique (X) qual(is):

Peso a mais	☐	Diabetes (tipo 1)	☐	Doença hepática	☐	Prefere não responder	☐
Obesidade	☐	Diabetes (tipo 2)	☐	Cancro	☐	Não aplicável	☐
Colesterol elevado	☐	Ansiedade	☐	Doença cardíaca	☐	Outra	☐
Pressão alta	☐	Depressão	☐	Stress	☐	________________	

9.7.2. Há histórico de doenças na família ascendente? Indique qual(is):

Peso a mais	☐	Diabetes (tipo 1)	☐	Doença hepática	☐	Prefere não responder	☐
Obesidade	☐	Diabetes (tipo 2)	☐	Cancro	☐	Não aplicável	☐
Colesterol elevado	☐	Ansiedade	☐	Doença cardíaca	☐	Outra	☐
Pressão alta	☐	Depressão	☐	Stress	☐	________________	

9.7.3. Há histórico de doenças na família descendente? Indique qual(is):

Peso a mais	☐	Diabetes (tipo 1)	☐	Doença hepática	☐	Prefere não responder	☐
Obesidade	☐	Diabetes (tipo 2)	☐	Cancro	☐	Não aplicável	☐
Colesterol elevado	☐	Ansiedade	☐	Doença cardíaca	☐	Outra	☐
Pressão alta	☐	Depressão	☐	Stress	☐	________________	

1. Há mais alguma coisa que gostasse de partilhar?

____

MUITO OBRIGADO PELA SUA COLABORAÇÃO.

This project has received funding from the European Union's Horizon Europe under grant agreement No 101080329

Funded by the European Union

Appendix C - Dataset 1 — Example raw parental questionnaire data

Table 7.1: Example data of parent–questionnaire responses (First two entries).

Variable / Question	ID 40	ID 74
Metadata		
Answer ID	40	74
Submission date	1980-01-01 00:00:00	1980-01-01 00:00:00
Last page	10	10
Initial Language	pt	pt
Seed	1249499194	199875597
Before you start, indicate the study code assigned to you, [ID:]	N001A	N001B
1. Physical activity and sport		
1.1. Do you practice any sports with regularity?	Yes	No
1.1.1. Yes answered: [How many times a week?]	1.0	-
1.1.1. Yes answered: [How many hours per session/training?]	0.75	-
1.1.2. If you answered yes, say the sport (s):	Yoga	-
1.2. Regarding physical activity, indicate the time in minutes (min) that you dedicate to: [Walk]	30 to 60 min per week	Less than 30 min per week
1.2. Regarding physical activity, indicate the time in minutes (min) that you dedicate to: [riding a bicycle]	Never	Less than 30 min per week
1.2. Regarding physical activity, indicate the time in minutes (min) that you dedicate to: [take care of the house]	More than 90 min per week	Less than 30 min per week
1.2. Regarding physical activity, indicate the time in minutes (min) that you dedicate to: [field activities]	Never	Less than 30 min per week
1.2. Regarding physical activity, indicate the time in minutes (min) that you dedicate to: [Activities garden or garden]	Less than 30 min per week	30 to 60 min per week
1.2. Regarding physical activity, indicate the time in minutes (min) that you dedicate to: [dance]	Less than 30 min per week	Less than 30 min per week
1.2. Regarding physical activity, indicate the time in minutes (min) that you dedicate to: [Other]	30 to 60 min per week	Less than 30 min per week
2. Digital environment at home		
2.1. What digital equipment do you have at home? [Television]	Yes	Yes
2.1. What digital equipment do you have at home? [Computer (fixed and/or portable)]	Yes	Yes
2.1. What digital equipment do you have at home? [Tablet]	No	No
2.1. What digital equipment do you have at home? [Mobile phone]	Yes	Yes
2.1. What digital equipment do you have at home? [Game console (GameBoy, Nintendo)]]	Yes	No
2.1. What digital equipment do you have at home? [Blu-ray/DVD/CD/Video Cassettes]	No	No
2.1. What digital equipment do you have at home? [Other]	-	-
2.1. Which ones do you usually use in the room? [Television]	No	No
2.1. Which ones do you usually use in the room? [Computer (fixed and/or portable)]	Yes	Yes
2.1. Which ones do you usually use in the room? [Tablet]	No	No
2.1. Which ones do you usually use in the room? [Mobile phone]	Yes	Yes

Variable / Question	ID 40	ID 74
2.1. Which ones do you usually use in the room? [Game console (GameBoy, Nintendo)]]	No	No
2.1. Which ones do you usually use in the room? [Blu-ray/DVD/CD/Video Cassettes]	No	No
2.1. Which ones do you usually use in the room? [Other]	-	-
2.2. Do you have internet at home?	Yes	Yes
2.2. If so, who usually uses the internet?	All	All
2.3. Regarding digital equipment, indicate what age was for the first time: 2.3.1. Television:	5.0	9.0
2.3.2. Computer:	12.0	18.0
2.3.3. Video game console:	-	-
2.3.4. Tablet:	30.0	-
2.3.5. Mobile phone:	18	20
2.4. Indicate the time of use of the following electronic devices: Computer (fixed and/or portable) [on a weekday (between Monday and Friday)]]	More than 4 hours	1 to 2 hours
2.4. Indicate the time of use of the following electronic devices: computer (fixed and/or portable) [on Saturday]	Up to 1 hour	Up to 1 hour
2.4. Indicate the time of use of the following electronic devices: computer (fixed and/or portable) [on Sunday]	Up to 1 hour	Up to 1 hour
2.4. Tablet [on a weekday (between Monday and Friday)]]	None	None
2.4. Tablet [on Saturday]	None	None
2.4. Tablet [on Sunday]	None	None
2.4. Mobile phone [on a weekday (between Monday and Friday)]	1 to 2 hours	3 to 4 hours
2.4. Mobile phone [on Saturday]	Up to 1 hour	3 to 4 hours
2.4. Mobile phone [on Sunday]	Up to 1 hour	3 to 4 hours
2.4. Television [on a weekday (between Monday and Friday)]]	None	Up to 1 hour
2.4. Television [on Saturday]	Up to 1 hour	1 to 2 hours
2.4. Television [on Sunday]	Up to 1 hour	1 to 2 hours
2.4. Video game console [on a weekday (Monday and Friday)]]	None	None
2.4. Video game console [on Saturday]	None	None
2.4. Video-game console [on Sunday]	None	None
3. Eating habits		
3.1. How many meals do you have a day?	4.0	3.0
3.1.1. Indicate the frequency of the following meals: [breakfast]	Every day	Every day
3.1.1. Indicate the frequency of the following meals: [morning snack]	Every day	2-4 days a week
3.1.1. Indicate the frequency of the following meals: [lunch]	Every day	Every day
3.1.1. Indicate the frequency of the following meals: [afternoon snack]	5-6 days a week	2-4 days a week
3.1.1. Indicate the frequency of the following meals: [dinner]	Every day	Every day
3.1.1. Indicate the frequency of the following meals: [supper]	1 day a week	Never or rarely
3.2. Indicate the place and frequency of main meals (breakfast, lunch, or dinner): [cooked/prepared at home]	Every day	5-6 days a week
3.2. Indicate the place and frequency of the main meals (breakfast, lunch, or dinner): [P.E., traditional Portuguese, Brazilian, Asian, African ...)]]]]	1 day a week	1-3 times a month
3.2. Indicate the place and frequency of the main meals (breakfast, lunch, or dinner): [fast food restaurants (e., Taco Bell McDonald's, pizza hut, burger king, kfc ...)]]]	1-3 times a month	1-3 times a month
3.2. Indicate the place and frequency of the main meals (breakfast, lunch, or dinner): [canteens/cafeterias (eg, schools)]]	Never or rarely	Never or rarely
3.2. Indicate the location and frequency of main meals (breakfast, lunch, or dinner): [pastries/cafeterias]	1-3 times a month	Never or rarely
3.2. Indicate the location and frequency of the main meals (breakfast, lunch, or dinner): [requests for application, Take Away, on site (eg, Uber, Bolt Glovo)]]]	1-3 times a month	1-3 times a month
3.2.1. Is it habit cooking? How often?	7.0	1.0
3.3.Indicate how often you consume the following types of food? [Cakes/pastries (e., donuts, bollycao)]	1 time a day	1-3 times a month
3.3.Indicate how often you consume the following types of food? [Breakfast cereals (e., Corn Flakes, Golden Grahams, Nesquik, Kellog's, Alpen)]	Never or rarely	1 day a week

Variable / Question	ID 40	ID 74
3.3.Indicate how often you consume the following types of food? [Cookies and cookies]	1-3 times a month	1 day a week
3.3.Indicate how often you consume the following types of food? [Chocolates, Gums and Subquited]	1 day a week	2-4 days a week
3.3.Indicate how often you consume the following types of food? [French potatoes (package), snacks, salty, popcorn]	1 day a week	2-4 days a week
3.3.Indicate how often you consume the following types of food? [Soft drinks or sugary drinks, packet/bottle juices (eg, fanta, coca-cola, compal)]]	5-6 days a week	2-4 days a week
3.3.Indicate how often you consume the following types of food? [Alcoholic beverages (such as beer, wine)]	5-6 days a week	More than 1 time a day
3.3.Indicate how often you consume the following types of food? [Fast-food (such as pizza, hamburgers, chicken nuggets)]	1 day a week	1-3 times a month
3.3.Indicate how often you consume the following types of food? [Processed (such as ham, sausages, bacon, ham)]]	1 day a week	2-4 days a week
3.4.Indicate how often you consume the following types of food? [Bread, cereals and tubers (such as potatoes), rice or pasta]	More than 1 time a day	More than 1 time a day
3.4.Indicate how often you consume the following types of food? [Dairy products (white milk, cheese, yogurts)]]	More than 1 time a day	1 time a day
3.4.Indicate how often you consume the following types of food? [Vegetable fat (oil, margarine)]	1 time a day	1 time a day
3.4.Indicate how often you consume the following types of food? [Vegetable fat (olive oil)]	1 time a day	1 time a day
3.4.Indicate how often you consume the following types of food? [Animal fat (butter, lard)]]	Never or rarely	Never or rarely
3.4.Indicate how often you consume the following types of food? [Red Meats (cow, pig, rabbit)]]	2-4 days a week	5-6 days a week
3.4.Indicate how often you consume the following types of food? [Poultry meats (such as chicken, turkey ...)]]	5-6 days a week	5-6 days a week
3.4.Indicate how often you consume the following types of food? [Fish]	2-4 days a week	2-4 days a week
3.4.Indicate how often you consume the following types of food? [Eggs]	2-4 days a week	5-6 days a week
3.4.Indicate how often you consume the following types of food? [Oilseed fruits (walnuts, almonds, hazelnuts, peanuts)]]	2-4 days a week	2-4 days a week
3.4.Indicate how often you consume the following types of food? [Legumes (peas, beans, lentils, beans)]]	1 day a week	2-4 days a week
3.4.Indicate how often you consume the following types of food? [Fruit (such as orange, apple, pear, banana)]]	More than 1 time a day	1 time a day
3.4.Indicate how often you consume the following types of food? [Wild mushrooms]	Never or rarely	Never or rarely
3.4.Indicate how often you consume the following types of food? [Cultivated Mushrooms]	1 day a week	1 day a week
3.4.Indicate how often you consume the following types of food? [Raw vegetables / salads]	5-6 days a week	5-6 days a week
3.4.Indicate how often you consume the following types of food? [Boiled vegetables]	5-6 days a week	5-6 days a week
3.4.Indicate how often you consume the following types of food? [Soup]	2-4 days a week	2-4 days a week
3.5. Do you have the habit of eating seasonal foods?	Yes	Yes
3.5.1. If so, which are the most consumed:	Fruits and Vegetables	Fruits and Vegetables
4. Food access		
4.1. Indicate how often you buy your food: [hypermarket]	1-3 times a month	1-3 times a month
4.1. Indicate how often you buy your food: [supermarket, mini -market, grocery store]	2-3 times a week	1 time per week
4.1. Indicate how often you buy your food: [market]	1-3 times a month	1 time per week
4.1. Indicate how often you buy your food: [Other]	1-3 times a month	1-3 times a month

Variable / Question	ID 40	ID 74
4.1.1. (Based on the results of the previous question) Why do you prefer to buy food in the (place you indicated)?	Supermarket	For the supply and ratio quality/price
4.2. Indicate the importance when choosing the place where you buy food (1 the least important and 6 the most important): [distance]	4	5
4.2. Indicate the importance when choosing where you buy food (1 the least important and 6 the most important): [price]	5	6
4.2. Indicate the importance when choosing where you buy food (1 the least important and 6 the most important): [product quality]	6	6
4.2. Indicate the importance when choosing the place where you buy food (1 the least important and 6 the most important): [other]	3	5
4.3. Indicate the importance of choosing the food you buy (1 the least important and 6 the most important): [Quality]	6	6
4.3. Indicate the importance of choosing the food you buy (1 the least important and 6 the most important): [ease of preparation]	6	4
4.3. Indicate the importance of choosing the food you buy (1 the least important and 6 the most important): [Freshness (not ultra-processed)]]]	5	5
4.3. Indicate the importance of choosing the food you buy (1 the least important and 6 the most important): [other]	1	3
4.4. Regarding access to food, for each question, answer what best suits your situation. In the last 6 months… [1. Did you feel concerned that the food will end at home before you have money to buy more?]	I don't know	No
4.4. Regarding access to food, for each question, answer what best suits your situation. In the last 6 months… [2. The food in your home ended before having money to buy more?]	No	No
4.4. Regarding access to food, for each question, answer what best suits your situation. In the last 6 months… [3. Has your family been out of money enough to get a healthy and varied diet?]	No	No
4.4. Regarding access to food, for each question, answer what best suits your situation. In the last 6 months… [4. Your family had to consume only the food you still had at home for having no money to buy more?]	No	No
4.4. Regarding access to food, for each question, answer what best suits your situation. In the last 6 months… [5. Did you jump a meal for not having enough money to buy food?]	No	No
4.4. Regarding access to food, for each question, answer what best suits your situation. In the last 6 months… [6. Has anyone adult in your family ate less for not having enough money to buy food?]	No	No
4.4. Regarding access to food, for each question, answer what best suits your situation. In the last 6 months… [7. Somewhere adult in your family was hungry but didn't eat for lack of money to buy food?]	No	No
4.4. Regarding access to food, for each question, answer what best suits your situation. In the last 6 months… [8. Any adult in the family spent a whole day without eating, with just one meal throughout the day, for not having enough money to buy food?]	No	No
4.4. The children/adolescents of your family ... (households with children up to 18 years old) [9. They could not have a healthy and varied diet (i.e. vegetables, fish, fruit,…) for lack of money?]	No	No
4.4. The children/adolescents of your family ... (households with children up to 18 years old) [10. Did you have to consume only a few foods they still had at home for being out of money?]	No	No
4.4. Children/adolescents in your family ... (households with children up to 18 years old) [11. Did they eat less than they should have no money to buy food?]	No	No
4.4. The children/adolescents of your family ... (households with children up to 18 years old) [12. Did you have fewer food on meals for not having enough money to buy more?]	No	No
4.4. Children/adolescents in your family ... (households with children up to 18 years old) [13. Did you stop making a meal because there is not enough money to buy food?]	No	No

Variable / Question	ID 40	ID 74
4.4. The children/adolescents of your family ... (households with children up to 18 years old) [14. They felt hungry, but did not eat for lack of money to buy food?]	No	No
5. Sustainable Eating		
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 fully agree) [L1. I understand and know how to apply hygiene and food safety practices (eg, wash your hands well, have short and clean nails, use apron whenever necessary)]]]	7	7
5.1.1. Sustainable Food Indicate what best suits your situation (1 being fully disagree and 7 I fully agree) [L2. I can distinguish between healthy and unhealthy foods]	7	6
5.1.1. Sustainable Food Indicate what best suits your situation (1 being fully disagree and 7 I fully agree) [L3. I know where to find information about food and nutrition]	5	5
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 I fully agree) [L4. I know the nutritional value of food]	5	5
5.1.1. Sustainable Food Indicate what best suits your situation (1 being fully disagree and 7 I fully agree) [L5. I know where to acquire good food for human health and the environment]	5	6
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 I fully agree) [L6. I know where food comes from (I understand and know what food chains are)]]	6	7
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 fully agree) [L7. I know the impact of food consumption on ecological and environmental sustainability (eg, food of the time, local foods, local markets)]]	5	6
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 I fully agree) [L8. I know the impact of food production on biological diversity (eg, sustainable use of natural, terrestrial, marine and freshwater resources)]]]	5	6
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 I fully agree) [L9. I know the impact of food production on climate change (eg, liberation of greenhouse gases, pollution)]]]	4	7
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 I fully agree) [L10. I know what "sustainable food" means (food that is "friend" of human health, ecology and the environment)]]]	6	6
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 fully agree) [L11. I know what "animal welfare"] means	6	7
5.1.1. Sustainable Food Indicate what best suits your situation (1 being totally disagree and 7 I fully agree) [L12. I know what "one health" means (collaborative approach, aiming to achieve ideal health results, recognizing interconnection between people, animals, plants, fungi and their shared environment)]]]	5	6
5.1.2. Food skills indicate what best suits your situation (1 being totally disagree and 7 fully agree). [Ca1. I think about what to buy before buying]	7	5
5.1.2. Food skills indicate what best suits your situation (1 being totally disagree and 7 fully agree). [Ca2. I am able to store food properly and safely]	6	6
5.1.2. Food skills indicate what best suits your situation (1 being totally disagree and 7 fully agree). [Ca3. I am able to use common cooking utensils and appliances]]	6	7
5.1.2. Food skills indicate what best suits your situation (1 being totally disagree and 7 fully agree). [Ca4. I am able to prepare and cook foods with basic/available ingredients]	6	7
5.1.2. Food skills indicate what best suits your situation (1 being totally disagree and 7 fully agree). [Ca5. I am able to adapt recipes based on available foods]	6	7
5.1.2. Food skills indicate what best suits your situation (1 being totally disagree and 7 fully agree). [Ca6. I am able to improvise meals based on available foods (including leftovers)]	6	6
5.1.3. Action intent and action strategies indicate what best suits your situation (1 being totally disagree and 7 fully agree). [IA1. I intend to practice healthy and sustainable diet]	7	7

Variable / Question	ID 40	ID 74
5.1.3. Action intent and action strategies indicate what best suits your situation (1 being totally disagree and 7 fully agree). [IA2. I will encourage other people to adopt healthy and sustainable diets]	7	7
5.1.3. Action intent and action strategies indicate what best suits your situation (1 being totally disagree and 7 fully agree). [IA3. I will adjust and practice healthy and sustainable diet]	6	6
5.1.3. Action intent and action strategies indicate what best suits your situation (1 being totally disagree and 7 fully agree). [EA1. I make shopping thinking about my 'ecological footprint']	6	6
5.1.3. Action intent and action strategies indicate what best suits your situation (1 being totally disagree and 7 fully agree). [EA2. I encourage others to reduce food waste]	6	6
5.1.3. Action intent and action strategies indicate what best suits your situation (1 being totally disagree and 7 fully agree). [EA3. I encourage others to have a healthy and sustainable diet]	6	6
5.1.3. Action intent and action strategies indicate what best suits your situation (1 being totally disagree and 7 fully agree). [EA4. Support campaigns associated with sustainable food or zero food waste]	6	6
5.1.4. Attitudes indicate what best suits your situation (1 being totally disagree and 7 I fully agree). [AT1. I believe I can have a healthy and sustainable diet]	7	7
5.1.4. Attitudes indicate what best suits your situation (1 being totally disagree and 7 I fully agree). [At2. I believe that the collective effort of consumers contributes to healthy eating and ecological and environmental sustainability]	7	7
5.1.4. Attitudes indicate what best suits your situation (1 being totally disagree and 7 I fully agree). [At3. I believe that the long -term effort of industries will contribute to improving healthy eating and ecological and environmental sustainability]	5	7
5.1.4. Attitudes indicate what best suits your situation (1 being totally disagree and 7 I fully agree). [At4. I believe that policies and legislation encourage the adoption of healthy eating and ecological and environmental sustainability]	4	7
6. Nature and environmental contact		
6.1. Indicate the distance from your home to the nearest green space (consider urban parks, woods, riverside parks, public gardens and community gardens):	150	150
6.2. Indicate how it considers the environment in the area of your residence (area in which you move): [air quality (the presence or absence of fumes and bad smells associated with traffic, industries)]]]]	Reasonable	Bom
6.2. Indicate how you consider the environment in the area of your residence (area in which you move): [public leisure and recreational spaces (such as squares, skate parks, maintenance circuit)]]]	Reasonable	Bom
6.2. Indicate how you consider the environment in the area of your residence (area in which you move): [green spaces of contact with nature (such as urban parks, woods and public access forests, riverside parks, green corridors, public gardens, community gardens)]]]]]	Reasonable	Very good
6.2. Indicate how it considers the environment in the area of your residence (area in which you move): [Urban afforestation (consider the presence of trees on the streets and public spaces)]]	Reasonable	Bom
6.2. Indicate how you consider the environment in the area of your residence (area in which you move): [Urban Cleaning and Maintenance (consider garbage collection, equipment, such as banks, drinking fountains, toilet access, gardening)]]]	Reasonable	Bom
6.2. Indicate how you consider the environment in the area of your residence (area in which you move): [comfort and safety to walk (consider the pedestrian floor, the presence of trees)]]	Bom	Reasonable
6.3. Indicate how often you contact the nature and green spaces. [Park with playground activities (such as Games Field)]	1-3 times a month	1-3 times a month
6.3. Indicate how often you contact the nature and green spaces. [Park without playgrounds]	1-3 times a month	1-3 times a month
6.3. Indicate how often you contact the nature and green spaces. [Forest spaces, public access forests, green corridors]	1-3 times a month	1-3 times a month

Variable / Question	ID 40	ID 74
6.3. Indicate how often you contact the nature and green spaces. [Riverside parks]	1-3 times a month	1-3 times a month
6.3. Indicate how often you contact the nature and green spaces. [Public gardens]	1-3 times a month	1-3 times a month
6.3. Indicate how often you contact the nature and green spaces. [Community gardens]	Never or rarely	Never or rarely
6.3. Indicate how often you contact the nature and green spaces. [Other]	Never or rarely	1-3 times a month
7. Sleep Quality		
7.1. Habits at bedtime and sleep duration last month: [hours to which they lie down most of the time] [from Monday to Friday]	23:00:00	22:00:00
7.1. Habits at bedtime and sleep duration last month: [hours I lie down most of the time] [on the weekend]	00:00:00	00:00:00
7.1. Habits at bedtime and sleep duration, during the last month: [hours you get up most of the time] [from Monday to Friday]	06:30:00	07:00:00
7.1. Habits at bedtime and sleep duration last month: [hours you get up most of the time] [on the weekend]	08:00:00	10:00:00
7.1. [How long (in minutes) did it take you to fall asleep most of the time] [from Monday to Friday]	15.0	5.0
7.1. [How long (in minutes) did it take you to fall asleep most of the time] [on the weekend]	15.0	5.0
7.1. [How many hours of sleep a night slept (it may be different from the number of hours in bed)] [from Monday to Friday]	6.5	9.0
7.1. [How many hours of sleep a night slept (it may be different from the number of hours in bed)] [over the weekend]	8.0	9.0
7.2. During the past month, indicate how often you had trouble sleeping because of: [take more than 30 minutes to fall asleep]	Less than 1 time per week	Never
7.2. During the past month, indicate how often you had trouble sleeping because of: [Wake up in the middle of the night or in the morning very early]	Less than 1 time per week	Less than 1 time per week
7.2. During the past month, indicate how often you had trouble sleeping because of: [getting up to go to the bathroom]	1 or 2 times a week	Less than 1 time per week
7.2. During the past month, indicate how often you had trouble sleeping because of: [having difficulty breathing]	Never	Never
7.2. During the past month, indicate how often you had trouble sleeping because of: [coughing or resonating high]	Less than 1 time per week	Less than 1 time per week
7.2. During the past month, indicate how often you had trouble sleeping because of: [feel very cold]	Never	Never
7.2. During the past month, indicate how often you had trouble sleeping because of: [feel very hot]	Never	Never
7.2. During the past month, indicate how often you had trouble sleeping because of: [having bad dreams or nightmares]	Less than 1 time per week	Never
7.2. During the past month, indicate how often you had trouble sleeping because of: [another reason. During last month, how many times have you had problems to sleep for this reason]	Never	Never
7.3. During last month, indicate how it would classify the quality of your sleep:	Boa	Boa
7.4. Over the last month, indicate: [took some medicine to sleep prescribed by the doctor, or indicated by someone else (friend, family, pharmacist), or even their initiative]	3 times a week or more	Never
7.4. Over the past month, indicate: [had problems staying up during meals while driving, or while participating in some social activity]	Never	Never
7.4. Over the last month, indicate: [felt little desire or lack of enthusiasm to perform your daily activities]	Never	Never
8. Mental well-being		
8. Indicate how much each statement applied to you last month. [1. I had difficulty calming me]	Nothing is applied to me	Applied to me a few times
8. Indicate how much each statement applied to you last month. [2. I felt my mouth dry]	Applied to me a few times	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [3. I couldn't feel any good feeling]	Nothing is applied to me	Nothing is applied to me

Variable / Question	ID 40	ID 74
8. Indicate how much each statement applied to you last month. [4. I felt difficulty breathing]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [5. I had difficulty taking initiative to do things]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [6. I had a tendency to react too much in certain situations]	Nothing is applied to me	Applied to me a few times
8. Indicate how much each statement applied to you last month. [7. I felt tremors (eg, in the hands)]]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [8. I felt nervous	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [9. I worried about situations where I could panic and make ridiculous figure]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [10. I felt that I had nothing to expect from the future]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [11. I gave myself to get agitated]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [12. I had difficulty relaxing]	Applied to me a few times	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [13. I felt discouragement and sadness]	Nothing is applied to me	Applied to me a few times
8. Indicate how much each statement applied to you last month. [14. I was -lerant about anything that prevented me from finishing what I was doing]	Nothing is applied to me	Applied to me a few times
8. Indicate how much each statement applied to you last month. [15. I felt almost panicked]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [16. I was not able to have enthusiasm for anything]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [17. I felt that I didn't have much value as a person]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [18. I felt that it was sometimes sensitive]	Applied to me a few times	Applied to me a few times
8. Indicate how much each statement applied to you last month. [19. I felt changes in my heart without exercise]	Nothing is applied to me	Applied to me a few times
8. Indicate how much each statement applied to you last month. [20. I felt scared without having a good reason for this]	Nothing is applied to me	Nothing is applied to me
8. Indicate how much each statement applied to you last month. [21. I felt that life made no sense]	Nothing is applied to me	Nothing is applied to me
9. Sociodemographics and health status		
9.1. Sex:	Woman	Men
9.1.1. Date of birth (dd/mm/aaaa):	23/08/1979	25/09/1975
9.2. The following questions are about the weight and height. [9.2.1. How much do you measure? (m)]	Yes	Yes
9.2. The following questions are about the weight and height. [9.2.1. How much do you measure? (m)] [Comment]	1.65	1.7
9.2. The following questions are about the weight and height. [9.2.2. How much do you weigh? (Kg)]	Yes	Yes
9.2. The following questions are about the weight and height. [9.2.2. How much do you weigh? (Kg)] [Comment]	66	81
9.2. The following questions are about the weight and height. [9.2.3. I prefer not to answer]	No	No
9.2. The following questions are about the weight and height. [9.2.3. I prefer not to answer] [Comment]	-	-
9.3. The following questions are about the place of residence and people who share the house with you: 9.3.1. [Parish:]	São Martinho do Bispo	São Martinho do Bispo
9.3. The following questions are about the place of residence and people who share the house with you: 9.3.1. [County:]	Coimbra	Coimbra
9.3. The following questions are about the place of residence and people who share the house with you: 9.3.1. [District:]	Coimbra	Coimbra
9.3. The following questions are about the place of residence and people who share the house with you: 9.3.1. [POST CODE:]	3045	3030 - 045

Variable / Question	ID 40	ID 74
9.3.2. Indicate (number), how many people share the house with you and what age: [until 4 years]	0	0
9.3.2. Indicate (number), how many people share the house with you and what age: [from 5 to 17 years]	1	1
9.3.2. Indicate (number), how many people share the house with you and what age: [from 18 to 64 years]	2	3
9.3.2. Indicate (number), how many people share the house with you and what age: [more than 65 years]	0	0
9.4. Education, indicate the level of education you completed:	Secondary and Post-Secondary Education (12 years)/ Complementary Course of the Lycles	3rd Cycle of Basic Education (9 years)/General Course of Lycles
9.4. Education, indicate the level of education you completed: [Other]	-	-
9.5. Are you working right now? [Yes]	Yes	Yes
9.5. Are you working right now? [Yes] [Comment]	-	-
9.5. Are you working right now? [No. Unemployed since:]	No	No
9.5. Are you working right now? [No. Unemployed since:] [Comment]	-	-
9.5. Are you working right now? [Retiree. What is the last profession?]	No	No
9.5. Are you working right now? [Retiree. What is the last profession?] [Comment]	-	-
9.5.1. If you are working, what is your profession? Please say exactly what you do, for example, civil servant, specify your activity and category)	Banker	Construction worker
9.5.2. Indicate the degree of physical intensity of your profession:	Very sedentary work	Great physical effort
9.6. What is your usual means of transportation in your daily activities?	Car	Car
9.7. In general, I would say that your health is:	Reasonable	Very good
9.7.1. Have you been diagnosed with any of these conditions? Indicate which one (is): [More weight]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [Obesity]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [high cholesterol]	No	Yes
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [high blood pressure]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [Diabetes (Type 1)]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [Diabetes (Type 2)]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which one (is): [Anxiety]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [Depression]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [liver disease]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [cancer]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [heart disease]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [stress]	Yes	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which one (is): [prefers not to answer]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which one (is): [not applicable]	No	No
9.7.1. Have you been diagnosed with any of these conditions? Indicate which (is): [Other]	-	-

Variable / Question	ID 40	ID 74
9.7.2. Is there a history of diseases in the ascending family? Indicate which one (is): [More weight]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [Obesity]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [high cholesterol]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [high blood pressure]	No	Yes
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [Diabetes (Type 1)]	Yes	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [Diabetes (Type 2)]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which one (is): [Anxiety]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [Depression]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [liver disease]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [cancer]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [heart disease]	Yes	Yes
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [stress]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which one (is): [prefers not to answer]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which one (is): [not applicable]	No	No
9.7.2. Is there a history of diseases in the ascending family? Indicate which (is): [Other]	-	-
9.7.3. Is there a history of diseases in the descending family? Indicate which one (is): [More weight]	Yes	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [Obesity]	No	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [high cholesterol]	Yes	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [high blood pressure]	No	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [Diabetes (Type 1)]	Yes	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [Diabetes (Type 2)]	No	No
9.7.3. Is there a history of diseases in the descending family? Indicate which one (is): [Anxiety]	No	Yes
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [Depression]	No	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [liver disease]	No	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [cancer]	No	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [heart disease]	Yes	No
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [stress]	No	Yes
9.7.3. Is there a history of diseases in the descending family? Indicate which one (is): [prefers not to answer]	No	No
9.7.3. Is there a history of diseases in the descending family? Indicate which one (is): [not applicable]	No	No

Variable / Question	ID 40	ID 74
9.7.3. Is there a history of diseases in the descending family? Indicate which (is): [Other]	-	-

Appendix D - Dataset 2 — Example raw body composition data

Table 7.2: Example data of Body Composition (First two entries).

Variable / Measure	ID 1	ID 2
1. Name	-	-
2. ID	n001	n002
3. Height	124,4	120,7
4. Date of Birth	2017.03.18.	2017.10.01.
5. Gender	M	M
6. Age	6,8	6,3
7. Mobile Number	-	-
8. Phone Number	-	-
9. Zip Code	-	-
10. Address	-	-
11. E-mail	-	-
12. Date of Registration	2024.01.22.	2024.01.22.
13. Memo	-	-
14. Test Date / Time	2024.01.22. 19:30:19	2024.01.22. 19:06:30
15. Weight	25,4	24,7
16. Lower Limit (Weight Normal Range)	20,5	19,3
17. Upper Limit (Weight Normal Range)	27,7	26,1
18. TBW (Total Body Water)	15,7	15,8
19. Lower Limit (TBW Normal Range)	13,6	12,8
20. Upper Limit (TBW Normal Range)	16,6	15,6
21. ICW (Intracellular Water)	9,7	9,8
22. Lower Limit (ICW Normal Range)	8,5	7,9
23. Upper Limit (ICW Normal Range)	10,3	9,7
24. ECW (Extracellular Water)	6,0	6,0
25. Lower Limit (ECW Normal Range)	5,1	4,9
26. Upper Limit (ECW Normal Range)	6,3	5,9
27. Protein	4,2	4,2
28. Lower Limit (Protein Normal Range)	3,6	3,4
29. Upper Limit (Protein Normal Range)	4,4	4,2
30. Minerals	1,56	1,46
31. Lower Limit (Minerals Normal Range)	1,26	1,18
32. Upper Limit (Minerals Normal Range)	1,54	1,44
33. BFM (Body Fat Mass)	3,9	3,2
BFM(\34. Lower Limit (BFM Normal Range)	2,9	2,7
35. Upper Limit (BFM Normal Range)	5,8	5,5
36. SLM (Soft Lean Mass)	20,2	20,3
37. Lower Limit (SLM Normal Range)	17,5	16,4
38. Upper Limit (SLM Normal Range)	21,3	20,0
39. FFM (Fat Free Mass)	21,5	21,5
FFM(\40. Lower Limit (FFM Normal Range)	18,5	17,4
41. Upper Limit (FFM Normal Range)	22,6	21,2
42. SMM (Skeletal Muscle Mass)	10,7	10,8
43. Lower Limit (SMM Normal Range)	9,2	8,6

Variable / Measure	ID 1	ID 2
44. Upper Limit (SMM Normal Range)	11,2	10,4
45. BMI (Body Mass Index)	16,4	17,0
46. Lower Limit (BMI Normal Range)	13,2	13,2
47. Upper Limit (BMI Normal Range)	16,8	16,8
48. PBF (Percent Body Fat)	15,4	13,1
49. Lower Limit (PBF Normal Range)	10,0	10,0
50. Upper Limit (PBF Normal Range)	20,0	20,0
51. FFM of Right Arm	0,78	0,91
52. Lower Limit (FFM of Right Arm Normal Range)	0,65	0,62
53. Upper Limit (FFM of Right Arm Normal Range)	0,89	0,84
54. FFM\55. FFM of Left Arm	0,77	0,91
56. Lower Limit (FFM of Left Arm Normal Range)	0,65	0,62
57. Upper Limit (FFM of Left Arm Normal Range)	0,89	0,84
58. FFM\59. FFM of Trunk	8,7	9,1
60. Lower Limit (FFM of Trunk Normal Range)	7,7	7,2
61. Upper Limit (FFM of Trunk Normal Range)	9,4	8,8
62. FFM\63. FFM of Right Leg	2,66	2,79
64. Lower Limit (FFM of Right Leg Normal Range)	2,35	2,16
65. Upper Limit (FFM of Right Leg Normal Range)	2,87	2,64
66. FFM\67. FFM of Left Leg	2,67	2,76
68. Lower Limit (FFM of Left Leg Normal Range)	2,35	2,16
69. Upper Limit (FFM of Left Leg Normal Range)	2,87	2,64
70. FFM\71. TBW of Right Arm	0,60	0,71
72. Lower Limit (TBW of Right Arm Normal Range)	0,52	0,48
73. Upper Limit (TBW of Right Arm Normal Range)	0,70	0,66
74. TBW of Left Arm	0,60	0,71
75. Lower Limit (TBW of Left Arm Normal Range)	0,52	0,48
76. Upper Limit (TBW of Left Arm Normal Range)	0,70	0,66
77. TBW of Trunk	6,7	7,1
78. Lower Limit (TBW of Trunk Normal Range)	6,0	5,6
79. Upper Limit (TBW of Trunk Normal Range)	7,3	6,9
80. TBW of Right Leg	2,07	2,17
81. Lower Limit (TBW of Right Leg Normal Range)	1,84	1,69
82. Upper Limit (TBW of Right Leg Normal Range)	2,24	2,07
83. TBW of Left Leg	2,08	2,15
84. Lower Limit (TBW of Left Leg Normal Range)	1,84	1,69
85. Upper Limit (TBW of Left Leg Normal Range)	2,24	2,07
86. ICW of Right Arm	0,37	0,44
87. Lower Limit (ICW of Right Arm Normal Range)	0,32	0,29
88. Upper Limit (ICW of Right Arm Normal Range)	0,44	0,41
89. ICW of Left Arm	0,37	0,44
90. Lower Limit (ICW of Left Arm Normal Range)	0,32	0,29
91. Upper Limit (ICW of Left Arm Normal Range)	0,44	0,41
92. ICW of Trunk	4,1	4,4
93. Lower Limit (ICW of Trunk Normal Range)	3,7	3,4
94. Upper Limit (ICW of Trunk Normal Range)	4,5	4,3
95. ICW of Right Leg	1,28	1,34
96. Lower Limit (ICW of Right Leg Normal Range)	1,14	1,04
97. Upper Limit (ICW of Right Leg Normal Range)	1,38	1,28
98. ICW of Left Leg	1,29	1,33
99. Lower Limit (ICW of Left Leg Normal Range)	1,14	1,04
100. Upper Limit (ICW of Left Leg Normal Range)	1,38	1,28
101. ECW of Right Arm	0,23	0,27
102. Lower Limit (ECW of Right Arm Normal Range)	0,20	0,19
103. Upper Limit (ECW of Right Arm Normal Range)	0,26	0,25
104. ECW of Left Arm	0,23	0,27
105. Lower Limit (ECW of Left Arm Normal Range)	0,20	0,19
106. Upper Limit (ECW of Left Arm Normal Range)	0,26	0,25

Variable / Measure	ID 1	ID 2
107. ECW of Trunk	2,6	2,7
108. Lower Limit (ECW of Trunk Normal Range)	2,3	2,2
109. Upper Limit (ECW of Trunk Normal Range)	2,8	2,6
110. ECW of Right Leg	0,79	0,83
111. Lower Limit (ECW of Right Leg Normal Range)	0,70	0,65
112. Upper Limit (ECW of Right Leg Normal Range)	0,86	0,79
113. ECW of Left Leg	0,79	0,82
114. Lower Limit (ECW of Left Leg Normal Range)	0,70	0,65
115. Upper Limit (ECW of Left Leg Normal Range)	0,86	0,79
116. ECW/TBW	0,381	0,381
117. ECW/TBW of Right Arm	0,378	0,379
118. ECW/TBW of Left Arm	0,377	0,379
119. ECW/TBW of Trunk	0,381	0,382
120. ECW/TBW of Right Leg	0,382	0,380
121. ECW/TBW of Left Leg	0,382	0,381
122. BFM of Right Arm	0,3	0,2
123. BFM\124. BFM of Left Arm	0,3	0,3
125. BFM\126. BFM of Trunk	0,6	0,1
127. BFM\128. BFM of Right Leg	1,0	0,9
129. BFM\130. BFM of Left Leg	1,0	0,9
131. BFM\132. InBody Score	81	82
133. Target Weight	25,3	24,7
134. Weight Control	-0,1	0,0
135. BFM Control	-0,1	0,0
136. FFM Control	0,0	0,0
137. BMR (Basal Metabolic Rate)	834	834
138. WHR (Waist-Hip Ratio)	0,73	0,73
139. Lower Limit (WHR Normal Range)	0,80	0,80
140. Upper Limit (WHR Normal Range)	0,90	0,90
141. VFL (Visceral Fat Level)	Level 1	Level 1
142. VFA (Visceral Fat Area)	13,5	9,3
143. Obesity Degree	105	109
144. Lower Limit (Obesity Degree Normal Range)	90	90
145. Upper Limit (Obesity Degree Normal Range)	110	110
146. BCM (Body Cell Mass)	13,9	14,0
147. Lower Limit (BCM Normal Range)	12,1	11,3
148. Upper Limit (BCM Normal Range)	14,7	13,9
149. AC (Arm Circumference)	20,5	20,9
150. AMC (Arm Muscle Circumference)	17,2	17,5
151. BMC (Bone Mineral Content)	1,32	1,18
152. Lower Limit (BMC Normal Range)	1,03	0,97
153. Upper Limit (BMC Normal Range)	1,26	1,19
154. TBW/FFM	73,1	73,6
155. FFMI (Fat Free Mass Index)	13,9	14,8
156. FMI (Fat Mass Index)	2,5	2,2
157. 1kHz-RA Impedance	494,0	420,0
158. 1kHz-LA Impedance	491,1	418,9
159. 1kHz-TR Impedance	29,6	28,6
160. 1kHz-RL Impedance	316,9	284,3
161. 1kHz-LL Impedance	313,7	291,8
162. 5kHz-RA Impedance	484,4	415,6
163. 5kHz-LA Impedance	483,2	413,4
164. 5kHz-TR Impedance	29,4	27,6
165. 5kHz-RL Impedance	311,4	281,1
166. 5kHz-LL Impedance	308,3	286,0
167. 50kHz-RA Impedance	442,0	384,3
168. 50kHz-LA Impedance	444,6	383,8
169. 50kHz-TR Impedance	25,6	24,6

Variable / Measure	ID 1	ID 2
170. 50kHz-RL Impedance	283,6	259,1
171. 50kHz-LL Impedance	281,1	263,4
172. 250kHz-RA Impedance	401,1	345,0
173. 250kHz-LA Impedance	402,5	345,8
174. 250kHz-TR Impedance	21,0	20,8
175. 250kHz-RL Impedance	254,8	230,8
176. 250kHz-LL Impedance	252,3	234,7
177. 500kHz-RA Impedance	385,0	330,6
178. 500kHz-LA Impedance	386,5	331,2
179. 500kHz-TR Impedance	18,7	18,9
180. 500kHz-RL Impedance	248,7	222,6
181. 500kHz-LL Impedance	245,6	226,5
182. 1000kHz-RA Impedance	371,2	318,6
183. 1000kHz-LA Impedance	372,3	319,0
184. 1000kHz-TR Impedance	13,6	16,0
185. 1000kHz-RL Impedance	244,5	218,3
186. 1000kHz-LL Impedance	243,7	222,0
187. 5kHz-RA Reactance	16,9	12,3
188. 5kHz-LA Reactance	15,2	11,5
189. 5kHz-TR Reactance	2,3	1,6
190. 5kHz-RL Reactance	10,9	8,3
191. 5kHz-LL Reactance	10,8	8,5
192. 50kHz-RA Reactance	35,4	31,5
193. 50kHz-LA Reactance	34,9	30,1
194. 50kHz-TR Reactance	4,8	3,2
195. 50kHz-RL Reactance	27,7	24,4
196. 50kHz-LL Reactance	26,5	24,3
197. 250kHz-RA Reactance	30,8	28,9
198. 250kHz-LA Reactance	31,6	28,3
199. 250kHz-TR Reactance	7,5	4,0
200. 250kHz-RL Reactance	26,2	22,9
201. 250kHz-LL Reactance	24,2	22,9
202. 50kHz-RA Phase Angle	4,6	4,7
203. 50kHz-LA Phase Angle	4,5	4,5
204. 50kHz-TR Phase Angle	10,8	7,5
205. 50kHz-RL Phase Angle	5,6	5,4
206. 50kHz-LL Phase Angle	5,4	5,3
207. 50kHz-Whole Body Phase Angle	5,2	5,1
208. Measured Circumference of Neck	24,1	25,4
209. Measured Circumference of Chest	64,8	65,7
210. Measured Circumference of Abdomen	54,4	54,3
211. Measured Circumference of Hip	74,9	74,8
212. Measured Circumference of Right Arm	20,5	20,9
213. Measured Circumference of Left Arm	20,5	20,9
214. Measured Circumference of Right Thigh	39,2	38,9
215. Measured Circumference of Left Thigh	39,2	38,9
216. Growth Score	108	107
217. Obesity Degree of a Child	100	107
218. Lower Limit (Obesity Degree of a Child Normal Range)	90	90
219. Upper Limit (Obesity Degree of a Child Normal Range)	110	110
220. Systolic	-	-
221. Diastolic	-	-
222. Pulse	-	-
223. Mean Artery Pressure	-	-
224. Pulse Pressure	-	-
225. Rate Pressure Product	-	-

Variable / Measure	ID 1	ID 2
226. InBody Type	770	770
227. Local ID	1803	1798
228. SMI (Skeletal Muscle Index)	0,0	0,0
229. Medical History	-	-
230. Group	-	-
231. Recommended Calorie Intake	1723	1695
232. Lower Limit (BMR Normal Range)	790	778
233. Upper Limit (BMR Normal Range)	883	869
234. SMM/WT	0,0	0,0

Appendix E - Dataset 3 — Example raw cardiorespiratory data

Table 7.3: Example data of Cardiorespiratory assessment (First two entries).

Variable / Measure	ID 1	ID 2
id	N001	N002
VO2 MAX (ML) (TP 1)	960	1078
VO2 max (ml/kg/min) (TP 1)	38.4	43.3
Power (Watts) (TP 1)	70	60
TIME (SEC) TP 1	450	420

Appendix F - Dataset 4 — Example raw functional data

Table 7.4: Example data of Functional Assessment (First two entries).

Variable / Measure	**ID 1**	**ID 2**
id	N001	N002
SEX	M	M
DATE OF BIRTH	3/18/2017	10/1/2017
DATE OF VISIT (TP1)	1/22/2024	1/22/2024
AGE (TP1)	6.85	6.31
HEIGHT (CM) (TP1)	124.4	120.7
WEIGHT (TP1)	25.4	24.7
SITTING HEIGHT (CM) (TP1)	68	65.75
20M SPRINT (TP1)	4.17	4.23
RIGHT HAND GRIP (TP1)	11.5	15
LEFT HAND GRIP (TP1)	12	13.5
SQUAT JUMP (TP1)	8.7	19.6
COUNTER MOVEMENT JUMP (TP1)	10.2	15.3
SEAT AND REACH (cm) (TP1)	32	29

TP1 means Time Point T1

Appendix G - Dataset 5 — Example raw drawings data

Table 7.5: Example raw coding from the Drawings dataset (First two entries).

Variable / Measure	**ID 1**	**ID 2**
1. What is your favourite food?		
Water	0	0
Soup	0	0
Vegetables	0	0
Fruits	0	0
Cereals & derivatives	0	0
Meat, fish & eggs	0	0
Ultra-processed food	1	1
2. What makes you grow strong and healthy?		
Water	0	0
Soup	0	0
Vegetables	1	0
Fruits	1	0
Cereals & derivatives	0	0
Meat, fish & eggs	0	1
Exercise	0	0
Nature/Environment	0	0
People	0	0
3. What do you enjoy doing the most?		
To play	1	0
Screen use related activities	1	0
Exercise related activities	0	0
To paint / To draw	0	0
To write / To calculate	0	0
To dance	0	0
To cook	0	0
Animals related activities	0	0
Activities with family and friends	0	0
Other	0	1

Variable / Measure	ID 1	ID 2
4. What is the best 'Superhero' to represent the PAS GRAS project? Main character		
Anthropomorphisms (fruits and vegetables)	0	0
Other anthropomorphisms	0	0
Human characters	1	1
Animal (cats)	0	0
Fruits	0	0
Scale	0	0
Vegetables	0	0
Main character ornaments		
Heart	0	1
Super powers	1	0
Animals	0	0
Bars and weights	0	0

Appendix H - Detailed index formulas

This appendix provides detailed, step-by-step procedures for computing the eight composite indices derived from the GicO parental questionnaire.

7.1.1 PAI – Physical Activity Index

- **Data used:** Section 1.1 and 1.2 (sports participation and physical activity categories)
- **Procedure:**
  1. Convert frequency × duration of sport sessions into weekly minutes
  2. Map daily activity categories (e.g., "30–60 min/week") to midpoints (e.g., 45 min)
  3. Assign MET values per activity type (e.g., walking = 3.3, cycling = 4.0)
  4. Convert all activity time into MET-h/week using:

$$\text{MET-h/week} = \frac{\text{min}}{60} \times \text{MET}$$

  5. Sum all MET-h/week and rescale to 0–100

7.1.2 STI – Screen Time / Sedentary Time Index

- **Data used:** Section 2.4 – screen time per device (weekday/weekend)
- **Procedure:**
  1. Convert ordinal categories into midpoint values (e.g., "1–2 h/day" = 1.5 h)
  2. Compute weighted average:

$$\text{Total screen time} = \frac{5 \times \text{weekday} + 2 \times \text{weekend}}{7}$$

  3. Linearly rescale to 0–100 scale (e.g., 0 = no screen, 6+ hours = 100)

7.1.3 DQI – Diet Quality Index

- **Data used:** Section 3 – food frequency questionnaire (32 items)

- **Procedure:**
  1. Map categories to portions/week (e.g., “2–3x/day” = 17.5/week)
  2. Group into healthy (e.g., vegetables, legumes) and unhealthy (e.g., sugary snacks) foods
  3. Assign weights: healthy items contribute positively; unhealthy negatively
  4. Sum scores, rescale to 0–100

7.1.4 FIS – Food Insecurity Score

- **Data used:** Section 4.4 – 14 binary items
- **Procedure:**
  1. Count number of “Yes” responses
  2. Final score ranges from 0 (no insecurity) to 14 (severe)

7.1.5 SFL – Sustainable Food Literacy

- **Data used:** Section 5 – 21 Likert items (L1–L12, CA1–CA6, IA/EA/AT)
- **SFL** – Sustainable Food Literacy: aggregates 21 Likert-scale items reflecting food knowledge, skills, and attitudes. These items were structured around three domains (knowledge, competencies, and behaviours), in alignment with the EU Food Literacy Wheel [**?**], which defines core dimensions of food literacy for public health promotion.
- **Procedure:**
  1. Reverse-code negatively worded items (if any)
  2. Sum all items
  3. Rescale to 0–100 scale

7.1.6 GAI – Green Access Index

- **Data used:** Section 6 – distance, perceived quality, usage frequency
- **Procedure:**
  1. Rescale:
     - Distance (metres): invert and rescale (e.g., 1500m = 0, 0m = 10)
     - Quality (0–5): scale to 0–10
     - Frequency: days/week, scaled to 0–10
  2. Compute average of the three scores and rescale to 0–100

7.1.7 SQI – Sleep Quality Index

- **Data used:** Section 7 – 19 items scored according to PSQI-PT
- **Procedure:**
  1. Follow PSQI scoring algorithm (7 components: 0–3 each)
  2. Sum components: total score 0–21

7.1.8 SAD – Stress, Anxiety, Depression

- **Data used:** Section 8 – 21 items from DASS-21
- **Procedure:**
  1. Group items into subscales:
     - Depression: items 3, 5, 10, 13, 16, 17, 21
     - Anxiety: items 2, 4, 7, 9, 15, 19, 20
     - Stress: items 1, 6, 8, 11, 12, 14, 18
  2. Sum raw scores per subscale and multiply by 2

7.1.9 SES- Socioeconomic Status Index

The Socioeconomic Status (SES) Index was computed using a combination of parental education, profession, and family dependency structure. The method was implemented as follows:

1. **Education Level (educ_num)**: The highest completed level of parental education was pre-coded on a scale from 0 to 8 and rescaled to a 0–5 range to align with the profession score.
2. **Profession Score (prof_num)**: Free-text responses regarding current or last held profession (if retired) were matched against a list of regular expressions corresponding to typical occupational prestige levels. For example, "doctor" and "engineer" received high scores (8), while "labourer" or "student" received lower values (1–3). If no match was found, a default score of 5.0 was assigned.
3. **Number of Dependent Children (n_child)**: A family dependency factor was derived by summing the number of children aged 0–4 and 5–17 residing in the household. Missing values were interpreted as zero.
4. **Handling of Missing Data**: Where education or profession scores were missing, they were imputed using the median value of the corresponding variable across the dataset.

5. **PCA-Based Scoring**: The education and profession scores were standardised and submitted to Principal Component Analysis (PCA) with a single component, which served as the base score representing socioeconomic status.

6. **Adjustment for Family Dependency**: The PCA-derived score was divided by $(1 + \text{n\_child})$ to account for the economic impact of having dependent children in the household.

7. **Normalisation and Categorisation**: The adjusted SES score was min–max normalised to a 0–100 scale. Subsequently, the values were categorised into four ordinal groups according to predefined thresholds:

$$\text{SES Category} \in \{\text{Low}, \text{Lower-Middle}, \text{Upper-Middle}, \text{High}\}$$

with thresholds defined as $(-0.1, 25, 50, 75, 100]$.

Appendix I - Portuguese Directorate-General of Health BMI curves

The following appendix includes the child-specific reference BMI curves provided by the Portuguese Directorate-General of Health [23].

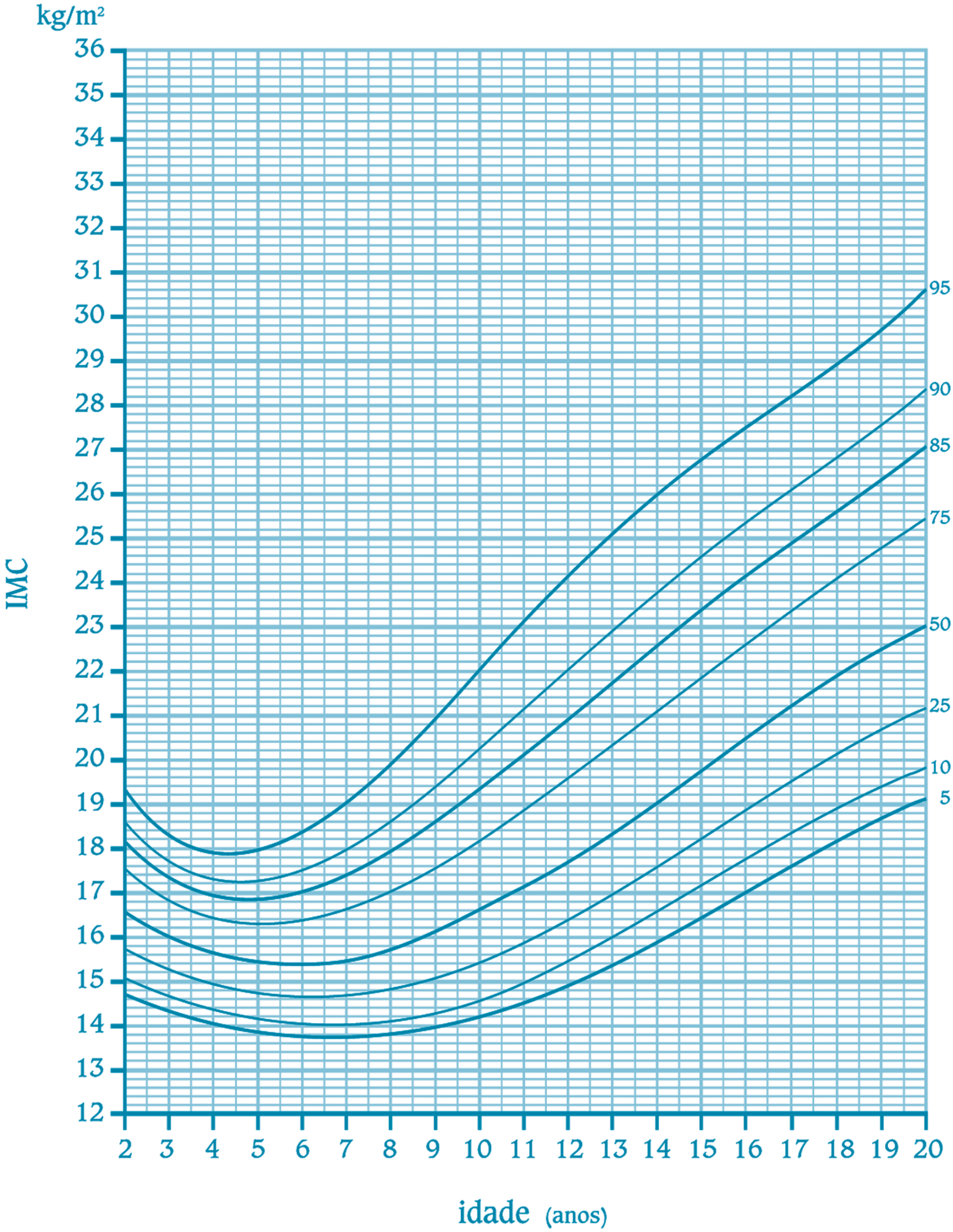
RAPARIGAS

índice de massa corporal 2-20 anos

OBESIDADE > percentil 95
EXCESSO DE PESO > percentil 85 e < percentil 95

RAPAZES

índice de massa corporal 2-20 anos



OBESIDADE > percentil 95
EXCESSO DE PESO > percentil 85 e < percentil 95

Instituto Superior
de Engenharia
Politécnico de Coimbra