Machine Learning Approach to Predict Childhood Neurodevelopmental Outcomes in the FinnBrain Birth Cohort Study: Importance of Serum Biomarkers

University of Turku
Department of Computing
Master of Science (Tech) Thesis, Digital Health
2024
Riikka Lund

Supervisors:
Prof. Antti Airola
Prof. Hasse Karlsson
Prof. Linnea Karlsson

UNIVERSITY OF TURKU
Department of Computing

Riikka Lund: Machine Learning Approach to Predict Childhood Neurodevelopmental Outcomes in the FinnBrain Birth Cohort Study: Importance of Serum Biomarkers

Master of Science (Tech) Thesis, Digital Health, 98 p., 4 app. p.

2024

The aim of this study was to determine whether serum biomarkers predict behavioural and socio-emotional problems of the children participating in the FinnBrain Birth Cohort study. The biomarkers were measured from maternal serum during pregnancy and children's own serum at five year follow-up. In addition, the aim was to identify other factors that may co-influence the outcomes. The outcomes of interest included Brief Infant Toddler Social Emotional Assessment Problem and Competence scores from two year follow-up and Strengths and Difficulties total difficulties scores from four and five year follow-ups.

The original data contained 6051 features and 1642 observations, including a panel of 13 biomarkers. After exploration and cleaning, the data was splitted into training and test datasets. The machine learning model was developed using training data and five-fold grid search cross-validation approach. The key steps included comparison and tuning of regressors and classifiers as well as techniques to mitigate class imbalance. The generalisation performances were evaluated in the hold-out test dataset and features predicting the outcomes were identified using permutation and SHAP techniques.

Acceptable performance levels were achieved using XGBoost Classifier and weighted target features for the models predicting total difficulties outcomes, however, not for Problem and Competence outcomes. The generalisation performances of the models on the holdout test data were moderate (ROC-AUC 0.63-0.66). Gestational TSH levels were among the most important features predicting total difficulties at both four and five year follow-ups. In addition, several other biomarkers, including LDL, APOA1, Trigly, FT4, Glucose HK2 and insulin, predicted the five year outcome with weaker influence. Furthermore, numerous other protective and risk factors were identified. Children's own biomarkers were not associated with the total difficulties. The results suggest that gestational imbalance in thyroid, lipid and glucose metabolism in combination with numerous other prenatal and early life factors influence the total difficulties outcome at five year follow-up.

This study is important in advancing our understanding of the early life factors associated with emotional and behavioural problems in the childhood and provide predictive markers for early detection of individuals at risk.

Keywords: gestational, serum, biomarker, BITSEA, SDQ, machine learning, XGBoost

Contents

1 Introduction

This project was part of the FinnBrain Birth Cohort Study. The main goal of the FinnBrain study is to advance our understanding of how prenatal (before birth) and early life environment and stress exposures may influence child's neurodevelopment and long-term health outcomes. Prenatal and early life stress exposure can have long-lasting impact on child's development and is know to be a risk factor for later behavioural and socio-emotional problems as well as several somatic and neuropsychiatric disorders, such as cardiovascular diseases and depression. In order to enable identification of biological and environmental factors mediating these associations, the prospective population based FinnBrain Birth Cohort Study was launched in 2010. Since then, rich longitudinal data, including biological samples and measurements from nearly 4000 families from pregnancy until childhood, has been collected [58]. This data resource at FinnBrain comprises thousands of variables evaluating early life environment of the participating children and aims to follow them until adulthood. The comprehensive dataset with prospective design facilitates identification biomarkers for stress exposures, developmental periods of vulnerability, as well as factors in the early life environment that either protect or increase the risk of adverse stress related health outcomes. [58].

The aim of this thesis project was to determine whether serum biomarkers collected during pregnancy and at five year follow-up predict or associate with the behavioural and socio-emotional problems of the children. To achieve this aim, supervised machine learning approach was applied. The biomarkers were measured from the maternal serum at gestational week 24 and from the children's own serum at the five year follow-up. As neurodevelopmental outcomes are influenced by complex interactions between biological and environmental factors, which are not well understood in this context, also other data available in the FinnBrain Birth Cohort study was included in the analysis. The behavioural and social-emotional outcomes examined in this study included Brief Infant-Toddler Social and Emotional Assessment (BITSEA) performed at the age of two years and Strengths and Difficulties Questionnaire (SDQ) performed at the ages of four and five years. To examine whether the serum biomarkers associate with these outcomes in the high dimensional FinnBrain dataset, I applied relevant guidelines from the Cross-industry standard process for data mining and machine learning approaches [27], and trained, selected and examined performance of machine learning models in solving the problem. Finally, the features best explaining the approved model and outcomes were identified.

2 Literature review

2.1 Neurodevelopmental and psychiatric problems

Clinically significant psychiatric problems are common in childhood. Among children under the age of five years, 8-10 % are affected by emotional, behavioral, and social problems, which are typically long-lasting and are associated with later risk of psychiatric disorders [45]. According to the report by et al. Piao 2022 on Global Burden of Disease study (1990-2019), 8.8 % of children and adolescents, at age group of 0-19 years, suffer from one or more of neurodevelopmental or psychiatric disorders. The most common disorders included anxiety (2.24 %), idiopathic developmental intellectual disability (1.96 %), attention deficit/hyperactivity disorder (ADHD, 1.85 %), depressive disorder (0.91 %), major depressive disorder (0.69 %), autism spectrum disorder (ASD, 0.42 %), dysthymia (0.23 %), bipolar disorder (0.18 %) and eating disorder (0.10 %). [86]

Leppänen et al. 2023 examined prevalence of neurodevelopmental and psychiatric disorders in 341,632 Finnish children born in 2001 - 2006 from birth until age of 12 years [68]. They found that 16.6 % of the children had at least one of these disorders. The disorders were more common in preterm (22.2 %) vs term (16.3

%) children ($p \leq 0.0001$). The mean age of the first diagnosis was seven years for the term children with a peak in incidence between four to five years. Most commonly the affected children had behavioural and emotional disorder (9.7 % of term children), such as ADHD, conduct disorder, social disorder, tics, or pervasive and specific developmental disorder (7.3 % of the term children), such as ASD.

Comorbidity of these disorders is common and phenotypes are heterogeneous. Boys are more often affected than girls. The disorders decrease the quality of life of both families and affected individuals and create economical burden in societies world-wide. To mitigate these issues identification of factors supporting wellbeing and enabling early intervention would be important. However, identification of the individuals with problems and tailoring the intervention according to personalised needs is a challenge. Therefore, more studies are needed to increase our understanding of the protective and risk factors as well as biological mechanisms influencing the neurodevelopmental and psychiatric health. [102, 82]

2.2 Normal brain development

2.2.1 First trimester

Human central nervous system (CNS) starts to develop between gestational weeks three and four. Before this stage a blastocyst has been formed from the fertilised egg. The inner cell mass containing pluripotent stem cells gives rise to three germ layers endoderm, ectoderm and mesoderm. The development of central nervous system starts with neurulation during which neural tube is from the ectoderm. Neurulation is followed by neurogenesis which begins at mid-first trimester. Dur-

ing this phase differentiated neurons develop from neural precursor cells. At the gestational week four the anterior neural tube starts to develop into different structures. The anterior part gives rise to forebrain, midbrain and hindbrain and the posterior part develops into spinal cord. By week seven all the brain regions present at birth have already evolved. The neurons start to migrate into their final circuits in the different brain regions and this process is continued until gestational week 26. By the week 11 cerebrum has been established and covers all the other regions of brain except medulla oblongata and cerebellum. [49, 69, 95, 97, 116]

2.2.2 Second trimester

During the second trimester, synapses which mediate the communication between neuronal cells start to develop. The process begins between gestational weeks 12 and 16 and continues throughout the life. Similarly to many other developmental processes of brain, synaptogenesis is dependent of lipids, particularly cholestrol. During the first weeks of development, the fetus relies on the maternal resources until glial cells develop and start to provide endogenous cholestrol in complex with ApoE chaperone. Glial cells including oligodendrocytes and astrocytes are produced from the same neural precursor cells as neurons. In addition, other types of glial cells, such as macrophage like microglia, which develop from the haematopoietic erythromyeloide precursors in the yolk sac, migrate and reside in the CNS parenchyma before development of blood-brain-barrier. Microglia are important source of cytokines in CNS and participate in several key processes in the developing brain, such as neurogenesis, synaptogenesis, apoptosis, pruning and formation of neural circuits. In addition to microglia, also three other types of macrophages

migrate to the central nervous system and reside in the interface of parenchyma and blood circulation. These are meningeal and perivascular macrophages, which similarly to microglia originate from yolk sac, and choroid plexus macrophages, which develop from both embryonic and adult haematopoietic cells. [8, 49, 53, 69, 97, 95, 116]

2.2.3 Third trimester and postnatal development

During the third trimester neurons become insulated by myelin. The main component of the myelin sheat is cholestrol. During the last four weeks of term pregnancy the density of dendrites and axons increases and the brain reach the maximum growth rate. After birth, depending on the neuronal activity, the connections between neurons start to gradually decrease. The connections that are often utilised are preserved. By the age of 2-3 years the brain has reached 90 - 95 % of the size of an adult brain. The synaptic density and myelination rate are at the maximum levels. Between the ages of 4 and 11 years the structures of the brain gradually mature and specialize. By the age of 20, the brain has reached the adult levels of synaptic density and neurotransmitters. The synaptic density and grey matter have decreased and myelination continues. The adult brain consists of about 86 x 10^9 neuronal and similar number of glial cells. These cells form a functional neural network, which determines responses to the signals received from body. Accurate timing and activity of this neural network is critical for the appropriate functioning of brain. [8, 49, 53, 69, 95, 97, 116]

2.3 Neurodevelopmental and psychiatric risk factors

2.3.1 Genetic factors and early environment

The prenatal period and the first years of life are critical for the development of brain and mental health. Several factors have been identified which can either protect from or increase risk of neurodevelopmental and psychiatric disorders. However, the causal mechanisms remain unclear. Both genetic and environmental components appear to be important. The genetic factors primarily involve the combined influence of several low penetrance alleles. Disorders caused by mutations in a single gene are rare. In addition to the personal genetic landscape, the developing brain is particularly sensitive to the environmental influences. During the fetal period availability and appropriate balance of nutrients, hormones, lipids, metabolites, immune factors and other components is essential for the normal development. Disturbances, such as imbalance in nutrients, maternal stress or anxiety, exposure to pollutants, substance use, endocrine or metabolic disorders and infections or immune system disorders can have long lasting impact on the health and increase risk of adverse outcomes. Furthermore, the development of brain continues after birth and the first years of postnatal life are important for the cognitive and socio-emotional development. Poor quality parental care and behaviour, socioeconomic environment, sleep, diet, exposure to abuse or interparental violence as well as cognitive ability in cross-talk with personal genetic factors can increase risk of emotional and behavioural disorders. In addition, factors such as premature birth and low birth weight can increase risk of adverse

outcomes. [49, 53, 80, 82, 95]

2.3.2 Hormonal and metabolic balance

Maternal thyroid dysfunction has been recognised as a risk factor for neurodevelopmental disorders in their offspring. Thyroid hormones are important for the developmental processes and metabolic regulation, including carbohydrate and lipid metabolism. Thyroid dysfunction is common in pregnant women. For example, in a Danish population based cohort study of 101,032 pregnancies approximately 4 % of the mothers were found to have a thyroid dysfunction before (2.0 %), during (0.1 %) or within 5 years (1.8 %) after pregnancy [5]. Importantly, extended study in the same population including 857,014 children reported that diagnosis and treatment of maternal hyperthyroid disorder after the birth was associated with increased risk of ADHD whereas hypothyroid disorder associated with increased risk of ASD in their children. Diagnosis made before the birth did not increase the risk of these developmental disorders in their children. [6]. Consistently, in a review by Fetene et al. 2017, maternal thyroxin levels during pregnancy were repeatedly associated with the ADHD symptoms of the children at the ages ranging from 3 to 10 years. The symptoms associated with both low and high hormone levels. Similarly symptoms of ASD at the ages of 3 and 6, were associated with maternal hypothyroidism. In contrast, there are also studies, which have not found association between symptoms of ASD and maternal thyroid hormone levels. One common of the limitation in these studies is that most of them measured the hormones at the beginning of second trimester, however, did not control the levels throughout the pregnancy. Furthermore, more studies are needed to understand

influence of confounding factors. [41]

Another metabolic condition that has been recognised as a risk factor for child's neurodevelopmental health is maternal gestational diabetes (GDM). Gestational diabetes is also common affecting approximately 7 % of pregnancies. While majority of the exposed children remain unaffected, also adverse associations with cognitive capacity, language development, attention, impulsivity and behaviour have been reported. [85]. For example, a Canadian study (Gen3G cohort) of 548 mother-child dyads found association between GDM (higher fasting glucose levels) and increased SDQ externalising scores at both 3 and 5 year follow-ups [38]. A British study (Avon Longitudinal Study of Parents and Children) consisting of 15,133 mother-child dyads reported association between higher maternal fasting glucose at first trimester of pregnancy and increased risk of conduct problems in their children at the ages of 4-16 years. In addition, higher maternal BMI during first trimester associated with increased hyperactivity problems. [63]

Similarly, emerging evidence suggests importance of lipid balance in neurodevelopmental health. For example, Avon Longitudinal Study of Parents and Children on 15,133 mother-child dyads found association of lower maternal high-density lipoprotein (HDL) levels during the first trimester of pregnancy with decreased hyperactvity problems whereas increased triglyserides at the second trimester were associated with increased hyperactivity problems. [63]. An American birth cohort study (Born in Bradford) of 1,369 children found association between lower levels of cord blood high-density lipoprotein (HDL), or higher levels of very-low-density lipoprotein or triglycerides, and risk of being evaluated as less competent in emotional and social skills at the five year follow-up. The assesement was performed by

teachers. [75]. In a Japanese study of 1199 children, higher maternal consumption of monosaturated fatty acids, α-linolenic acid, ω-6 polyunsaturated fatty acids, and linoleic acid during pregnancy was found to associate with childhood emotional problems. [77].

2.3.3 Inflammatory factors

Besides hormonal and metabolic factors, importance of inflammatory factors in modulating neurodevelopmental health has also been recognised. Although inflammatory factors are important in mediating in normal development and protecting cells and tissues from injuries and pathogens, disturbances in the balance, such as maternal infections during pregnancy can be detrimental to the early neuronal development. [49]. For example, a Swedish study examined maternal infections during pregnancy and later ASD diganosis in their children in a total sample of 2,371,403 individuals. They found association between any inpatient diagnosis of infection with increased risk of ASD (approximately 30 %). The risk was independent of the timing of infection during pregnancy. [66]. Similarly, a Danish Medical Birth Register study (N = 1,612,342) found that children of mothers who had viral infection during the first trimester, or bacterial infection the during second trimester of pregnancy, were more likely to have ASD diagnosis during the mean follow-up time of 15.1 years. The average age of diagnosis was 9.3 years. [7]. Another Danish birth cohort study (N = 1,206,600) found also association between maternal infections and disorders, such as anxiety, behavioral and emotional disorders and developmental disorders, including ASD, in their children. However, according to their results the risk was not limited to the period of pregnancy. They

found that the risk of any mental disorders was 9 % higher in children whose mothers had been treated with anti-infective drugs during second or third trimester and 21 % higher if hospitalisation was required during third trimester. The risk was elevated also for infections before or after pregnancy. Based on their results the origin of infection did not influence the outcome. A question was raised whether the association can be explained by shared genetic susceptibility between mental disorders and infections. [74].

In addition to clinical infections, several studies have examined association between inflammatory biomarkers and neurodevelopmental disorders. For example, a Danish COPSAC2010 cohort study examined 604 mother-child dyads and found association between higher maternal CRP levels (gestational week 24) and increased ADHD diagnosis at 10 years follow-up. [93]. In contrast, a Finnish study of 1079 cases and 1079 matched controls found no association between maternal CRP levels during pregnancy (gestational weeks 8–12) and ADHD diagnosis in children born in 1998-1999. [34]. Numerous studies have also reported association between levels of cytokines, such as increased levels of IL6 [51, 54], GMCSF, IFNG, IL1A [54], IL4, IL1B [60], IL8 [51] measured from child's own tissues with ASD.

Although the findings so far suggest a link between early inflammatory exposure and neurodevelopmental disorders, the results accross studies are not coherent and causative links remain to be established. [49].

2.4 The FinnBrain Birth Cohort Study

The goal of the FinnBrain Birth Cohort Study is to provide new information of the child's brain developmental trajectories, long-term health effects of prenatal exposures, and the relevant biological mechanisms. For this purpose, the Study has collected rich multimodal longitudinal data with thousands of variables, biological samples and measurements from nearly 4000 families from pregnancy until childhood [58]. The influence of metabolic and inflammatory factors on child outcomes has not been extensively studied yet. According to the recent findings maternal psychological symptoms [59] as well as tiredness [55] during pregnancy associate with altered plasma cytokine profiles. On the other hand maternal depressive symptoms during early pregnancy were found to associate with DNA methylation status of genes important for neurogenesis and neuronal differentiation in placenta and potential upstream regulators included both hormonal and inflammatory signaling cascades. [71]. Furthermore, maternal BMI during pregnancy was found to associate with infant's brain structures [91] and functional networks [89]. Given that the prenatal imbalances in metabolic, hormonal and inflammatory factors have been recognized as risk factors for adverse health outcomes, a key question that remains open is how these factors in combination with other exposures influence long-term outcomes of the FinnBrain children.

The FinnBrain Birth Cohort study has assessed neurodevelopmental outcomes and neuropsychiatric health of the children through various approaches. Among these are two questionnaire studies, Brief Infant Toddler Social Emotional Assessment (BITSEA) and Strengths and Difficulties Questionnaire (SDQ), which are used in this thesis work and therefore described in more detail in the following subsections.

2.4.1 Brief Infant Toddler Social Emotional Assessment

In the FinnBrain Birth Cohort study, BITSEA has been used to assess outcomes of the children at the two year follow-up. This questionnaire study has been found to be a reliable method for detection of social, emotional or behavioral problems in the early childhood and has potential for identification of children with psychopathology, such as autism spectrum disorder, disruptive behavior, anxiety or depression [20, 22, 57, 61]. It correlates well with another questionnaire, the Child Behavior Checklist (CBCL), which is also commonly used to screen behavioural problems in early childhood. [21, 48]. The questionnaire is designed for parents and child-care personnel for screening of children at the age of 12- to 36-months. It contains 42-items evaluating different subscales and is a shorter version of previously established 169-item ITSEA questionnaire. [21].

Of the 42 items, 6 screen for problems in externalising, such as over-activity or impulsivity, defiance or aggression and aggression with peers. Internalising is assessed with 8 items evaluating such as fear, nervousness, worry, anxiety and social withdrawal. Eight items assess dysregulation, such as negative emotionality, sleeping or eating problems and unusual sensory sensitivity. Competence is assessed with 11 items, evaluating empathy, attention, motivation, imitation or pretended play and prosocial interaction. Symptoms of autism spectrum disorder are screened with 17 items evaluating such as repetative actions, monotonous phrases and social competence. Of the items 14 screen for clinically significant problems, such as self-harm or unresponsiveness to pain, which potentially indicate psychopathology. Some of the items overlap in different sections. [2, 21, 22].

The sum score is calculated for each question from the responses with point scale ranging from 0 (not true / rarely), 1 (somewhat true / sometimes) to 2 (very often true / often). The total problem score ranges from 0 to 62 and competence score from 0 to 22. High total problem score and low total competence score indicate possibility of social–emotional and behavioral problems and / or deficits in the competence of the child. [2, 21, 22]. For the total problem score the original developers of the questionnaire study have suggested the use of 85th and 90th percentiles as cut offs indicative of potential problems. Also 75th percentile cut off has been used in previous studies and is the recommended cut off in the manual. For competence scale lower 15th percentile or below may indicate deficit or delay [3, 21, 22, 65].

Lavigne et al. 2016 reviewed classification accuracy of BITSEA in different studies. They found mean specificity of 0.80 and sensitivity of 0.82 with original cut offs based on five different studies [65]. However, the thresholds may vary in different settings and are influenced for example by the age, sex and population of origin [21, 48, 61, 57]. Haapsamo et al. 2009 compared BITSEA scores from a sample of 50 infants with an average age of 18-months from Northern Finland to those obtained in US population. They found that the mean competence scores in Finnish sample and US population were rather similar, 17.98 vs 17.5, respectively. However, the mean total problem score in Finnish sample was lower (7.21) than in the US sample (9.6) [48]. Some studies have determined the optimal cut-offs by estimating performance of the models when using different thresholds [61, 44]. The problem score has also been used as a continuous outcome variable [107].

2.4.2 Strengths and Difficulties Questionnaire

Strengths and Difficulties Questionnaire (SDQ) is another screening tool, which has been widely used to assess psychological wellbeing of children. Several versions of the SDQ has been developed for different purposes. The version for assessing children at the age of 4-17 years is designed to be filled by parents, teachers or by the youths themselves. All of the versions include five different scales each with 5 items assessing 1) emotional symptoms, 2) conduct problems, 3) hyperactivity/inattention, 4) peer relationship and 5) prosocial behaviour. For each item the answer can be 'Somewhat True', 'Not True' or 'Certainly True'. The scoring depends on the question and for each scale can range from 0 to 10, if all the items are answered. The scales 1-4 are combined to obtain a total difficulties score (range 0-40) assessed by 20 items. Externalising score can be obtained by summing of conduct and hyperactivity scales and internalising score by summing emotional and peer problem scales. Both can range from 0 to 20. The scores from scales 1-4 can also be used individually. [46, 96]

The scores can be utilised in several manner. They can be used as continuous scores or as categorised classes. For categorisation, thresholds have been established based on surveys on UK population. The older version includes thresholds for three ranks 'normal', 'borderline' and 'abnormal' and whereas in the newer version four ranks can be used: 'Close to average', 'Slightly raised (/slighlty lowered)', 'High (/low)' and 'Very high (/very low)'. Approximately 10 % in the UK population are estimated to belong to the rank abnormal and 10 % to the rank borderline. [46, 96]. The recommended thresholds for each rank depend on who has completed the questionnaire and it may be necessary to adjust the thresholds based on the ethnic

background, age, gender and purpose of the questionnaire [101]. For example, in the Nordic countries the mean scores have been found to differ from those in UK [81]. The cutpoints for different categories have been suggested to be 2-5 units lower in the Finnish population [17].

Stone et al. 2010 characterised the psychometric properties of SDQ, which had been used to assess 4-12 year old children in 48 different studies including a total of 131,223 participants. They found that in general the consistency and reliability of the results was satisfactory when the responses by the parents and teachers were compared. However, when individual subscales were evalutated, the questionnaires filled by the teachers were found to be more reliable. The results from SDQ were found to correlate sufficiently with the other assessments of psychopathology. [101]. Lavigne et al. 2016 reviewed reported classification accuracies of different behavioural screening tools, including SDQ. Based on their results from 19 studies, the mean sensitivity of SDQ total difficulties score was 0.65 and specificity was 0.76. The accuracy was dependent on the cut offs that were used in different studies. Often better performances were achieved, if higher cut offs than those recommended in the original study (=17) were used. Lavigne et al. also found that SDQ had better performance than CBCL in assessing school age children. [65]

2.5 Machine learning

Machine learning can be considered as a subfield of computer science and artificial intelligence, which extends classical statistics and uses mathematical algorithms to

mine new information and knowledge from the data. A wide variety of techniques have been developed for different tasks, such as visualization, pattern discovery, and for predicting or finding associations and explanations based on given datasets. Different machine learning approaches include such as unsupervised, supervised, semi-supervised and reinforcement learning. In unsupervised learning the outcome labels of the data are not known. A typical example of unsupervised tasks is clustering. Clustering methods, such as self organising maps, can be used to identify for example patterns or similar groups present in the input data. Unsupervised learning can also be used for dimensionality reduction to capture and represent the most important variation present in the data as aggregated features. Commonly used methods include, such as principal component analysis (PCA), which aims to capture the variance present in the data by finding linear combinations of the original features and transforming them into so called principal components. [15, 16, 36].

In supervised learning the outcome labels of groups of interest are known and are used to train the algorithm to make predictions. Common supervised tasks include regression, which can be used to predict continuous numerical outcomes and classification to predict categorical outcomes. Semi-supervised learning is similar to supervised learning, however, in this case only part of the outcome labels are known. In reinforcement learning the algorithm learns in interaction with the environment by maximising the cumulative rewards that are given to enforce the desired behaviour. Deep learning, which exploits artificial neural networks, is a current state-of-the-art approach and can be applied in supervised, semi-supervised or reinforcement settings for a wide variety of machine learning tasks. However,

this technique is not optimal for small datasets and the resulting models are not transparent or easy to interpret. [15, 16, 36].

2.5.1 Intelligent data analysis

Application of machine learning to solve different tasks is an iterative process, which typically requires interdisciplinary knowledge in data science and statistics as well as domain specific expertise. Data science project applying machine learning involves several steps aiming to understand, process, analyse and interpret the data and results from datasets with variable quantities, qualities and types. During the process several decisions need to be made how to process and analyse the data and which techniques to use among numerous available options. Therefore, to ensure high quality of the results, a well structured and systematic approach is important for managing the project throughout its entire life cycle. [15]

Several frameworks formalising such 'Intelligent data analysis' [50] strategy have been developed [15]. Among these is Cross-Industry Standard Process for Data Mining (CRISP-DM) [27], which has become a widely adapted model for data science projects [15]. Although, the process was originally developed for industry, the guidelines are applicable and provide good standards also in academic settings. CRISP-DM divides data analysis project into six phases (Table 2.1). The first phase is the 'Business or project understanding' aiming to understand and define the objectives and requirements of the project as data analysis problem. The second phase is 'Data understanding', which involves gathering and exploration of the data to identify any potential issues for example in the quality, suitability and sufficiency of the data to address the project objectives. The third phase is 'Data

preparation' during which the data is processed into format suitable for modelling.

The fourth phase of CRISP-DM is 'Modelling' during which the methods are selected, optimised and assessed for technical and generalisation performance. The fifth phase 'Evaluation' involves evaluation of the process and results in relation to the original project objectives. During this phase any issues or challenges are described, approved models are described and justified decisions of the next steps are made for example whether the project can proceed to deployment phase or whether iterative process should be continued for further improvements. In the sixth phase 'Deployment' the plan for deployment, monitoring and maintenance as well as final report is prepared and process is reviewed for lessons learned, to pinpoint any challenges and suggestions for further development. Each of the phases includes multiple general tasks, as well as more specialised subtasks. According to the model, the process is flexible and the order of tasks may deviate from the original guidelines. In addition, it is iterative meaning that the process may return back to the earlier phases to improve or correct the previously made assumptions and solutions. [27]

2.5.2 Modelling techniques

During data analysis project one of the key steps is selection of suitable technique for modelling. This step is important as it can influence the outcome and success of the entire project. Selection of the modelling technique, however, is not trivial as a large number of algorithms belonging to different families have been develop to solve different tasks and new extensions are continuously developed to improve performance of the existing ones. Importantly, performance of the algorithms

Table 2.1: Brief overview of Cross-Industry Standard Process for Data Mining (adapted from Chapman et al. 2000)

Tasks	Outputs
1. Project understanding	
Determine objectives	Description of the background and status, objectives and success criteria.
Evaluate situation	Description of resources, assumptions, requirements, limitations, benefits, risks etc.
Technical data science goals	Description of expected results and success criteria.
Project plan	Plan for achieving the goals, initial assessment of the tools and techniques.
2. Data understanding	
Collect data	Description of data, location and collection process.
Describe data characteristics	Data characteristics and compliance with requirements.
Explore data	Data exploration report with key data characteristics and and initial findings.
Verify quality	Data quality report with possible solutions to any problems.
3. Data preparation	
Select data	Describe selected data with inclusion/exclusion criteria.
Clean data	Cleaning report including taken actions.
Construct data	Description of any derived features, new records or transformations.
Integrate data	Merged data (from multiple sources).
Format data	Reformatted data (e.g. ordered, sorted, with syntatic modifications).
4. Modelling	
Select modelling technique	Selected techniques and possible assumptions made on the data.
Generate test design	Test design for training, validating, testing and evaluation of the models.
Build model	Models, interpretation, issues and chosen parameters with selection criteria
Assess models	Report of model performances and possible rank (iterate until no improvement).
5. Evaluation	
Evaluate results	Result evaluation considering objectives, approved models and any issues.
Review process	Process review report, issues and suggestions for further improvement.
Determine next steps	Justified next steps: continue improvement or proceed to deployment.
6. Deployment	
Plan deployment	Action plan for deployment strategy.
Plan monitoring and maintenance	Action plan for monitoring strategy.
Produce final report	Report of the project, deliverables and possible results.
Review project	Summary of experiences, issues, challenges and needs for further improvement.

is dataset specific and therefore typically several different algorithms need to be tested to find the optimal one. In addition, depending on the algorithm, variable numbers of hyperparameter may need to be tuned to optimise its performance. [36]

Only a few studies have performed comprehensive comparisons of machine learning algorithms to understand their performance across various types of datasets. Fernandez-Delgado et al. 2014 compared performance of 179 classifiers belonging to 17 families in predicting outcomes of 121 different types of datasets. The families of algorithms included Bayesian, boosting, bagging, decision trees, dis-

criminant analysis, generalized linear models, logistic and multinomial regression, nearest neighbors, neural networks, random forests, rule-based classifiers, support vector machines, stacking and other ensembles and several other methods. They found random forest and Support Vector Machine with Gaussian kernel to have the best performance accross datasets (both ≤ 90 % accuracy). Also neural networks and boosting ensembles were among the best performing classifiers. [40]. More recently, algorithms belonging to the newer generation of gradient boosting family, have been found to have at least equally good or better performance than random forest [117, 13] or support vector machines [117].

Also large scale studies comparing regressors are limited. Fernandez-Delgado et al. 2019 compared performance of 77 regressors across 83 datasets. [39]. They found cubist (M5 rule-based model with corrections based on nearest neighbors) and bstTree (the boosting ensemble of regression trees) to have the best performance across small and large as well as difficult and easy datasets. In addition, gradient boosted machine and M5 rule-based model had also good performance with the most of datasets. Although these studies provide good candidates to be evaluated for different datasets, how well the results generalise to other types of datasets, in particular to high-dimensional epidemiological data is less clear. For example, UCI repository used by Delgado et al. 2014 contains only nine Health and Medicine datasets with more than 100 features and only four of the datasets contain more than 1000 features [105]. In general, despite the methodological developments and increased sample sizes, the 'No Free Lunch Theorems' introduced by Wolpert and Macready in 1997, which states that 'for any algorithm, any elevated performance over one class of problems is offset by performance over another class' still holds

[112].

2.6 Machine learning in epidemiological and health research

Epidemiological research, such as the FinnBrain Birth Cohort Study, examines the health of populations in order to identify risk factors and causes that can compromise health. The purpose is to identify patterns in geographic regions and timing of the outcomes and to develop interventions for improving the health outcomes. The central methodological approach in epidemiological research has been traditional statistics. Statistical methods have key importance in determining sufficient sample sizes, description of data, and in examining relationships or differences between attributes of interest in hypothesis driven manner. [24] More recently use of machine learning techniques has started to emerge. Machine learning provides powerful alternative for epidemiological research with continuously increasing data sizes. When the number of variables in the input datasets increases, selection of the covariates, for example by using forward or backward selection, and sufficient control of confounders can become a challenge for the traditional statistical approaches. [64] In comparison to classical approaches, machine learning techniques can manage high dimensional data without prior knowledge of the feature importance and learn the model directly from the data. Another advantage is the data agnostic nature of the machine learning models. Whereas use of inferential statistical methods is typically based on strong assumptions of the data, such as normality, homoschedasticity, absence of multicolinearity and dependencies be-

tween variables, machine learning models are less sensitive and typically do not require such stringent assumptions. [15, 36]. However, as machine learning has not been yet used extensively in epidemiological research it is not well established which algorithms typically have the most optimal performance in addressing specific epidemiological questions and what is the sufficient sample size considering the features present in the data. [64]

2.6.1 Previous studies using machine learning in predicting health outcomes

Morgenstern et al. (2020) reviewed 231 studies which have used machine learning to predict health outcomes. The purpose of the review was to elucidate which outcomes have been examined, what are the most common data sources, and whether reporting follows the established guidelines. The reviewed studies were performed between 1980 and 2018 and the ones using logistic regression were excluded. According to their findings the most commonly studied health outcome was cardiovascular disease (n=22), followed by influenza (n=15), dengue fever (n=14), healthcare utilization (n=14), mortality (n=13), suicidality (n=13), cancer (n=12) and perinatal health (n=12). The median sample size in the studies was 5,414 and number of features 17. Most commonly the datasets were collected from health registers (n=126) or were generated by the researchers (n=86). The most popular algorithm was neural networks (n=95) followed by support vector machines (n=59), single tree based methods (n=52), and random forests (n=48). Typically machine leaning estimators were compared to traditional statistical models (n=111), such as logistic regression. The most common input features included

disease history, age, sex, smoking and meteorology. Most common validation approach was hold out (n=112). Only 15 studies performed external validation and 32 reported no validation approach. Most studies used area under the curve (AUC, n=98) to evaluate the model performance. Other common metrics were accuracy (n=76) and recall (n=68). The authors of the review stated that assessing performance of the model to predict probability of the outcome, model calibration, would be very important, however, was rarely performed. Only 77 studies reported overall performance of the model, typically with root mean square error (n=35). They also concluded that the adherence to the established guidelines for reporting the results was limited. [79]

Battineni et al. 2020 reviewed use of machine learning in diagnosis of chronic diseases. They found 453 studies published in 2015-2019, however, after filtering only 22 were left for in-depth analysis. According to their findings the most commonly used models included support vector machines (SVM, 45 %), K nearest neigbors (KNN) and Naive Bayes (23 %), Logistic regression (18 %) and random forest (14 %). They also examined accuracies of the models in predicting different diagnoses. They found that prediction of diabetes had accuracy of 73.1–91.6 %, cardiac diseases 84–91 %, liver diseases 78.1–82.7 %, depression 72-80 % and Alzheimer's disease 79 %. The accuracies were influenced by the algorithm, data type and input data. [12]

2.6.2 Previous studies using machine learning to predict BITSEA and SDQ outcomes

Only few studies exist which have applied supervised machine learning methods to predict BITSEA or SDQ outcomes. Usta et al. 2020 trained five different machine learning models, including Decision Trees, Linear model, Logistic Regression, Naive Bayes and SVM to predict outcomes of 2775 children assessed with BITSEA and ASQ questionnaires. In their approach, features with over 30 percent of missing values were excluded and the missing values for the remaining ones were imputed by using KNN method. They used principal compinent analysis (PCA) for dimensionality reduction and transformed outcome scores for classification by using cut-off values. Features with over 60 percent of correlation were excluded from the input data. The final dataset had 87 features and 53 output features. Ten fold cross-validation was used for evaluating the model performance based on AUC values. According to their results, SVM had the best accuracy (90.8 percentage) followed by Naive Bayes (81.3), Logistic regression (70.2), Linear model (68.3) and Decision tree (65.0). In the final model mother's lack of interest in things, trouble in concentrating, father's level of education, mother's feeling of people's unfriendliness or dislike, problems in falling asleep, father's health issues, duration of breastfeeding and unplanned pregnancy had the highest weights. [106]

Liverani et al. 2023 examined performance of the linear SVM Classifiers in predicting behavioural outcomes of 80 children born before gestational weeks 32 in Geneva and Lausanne. The input features included volumetric magnetic resonance imaging (MRI) data for brain structures measured at the age corresponding to 38.3–41.9 gestational weeks and several clinical prenatal and parental socioeconomic factors

(Largo scale). The outcomes of the children were assessed with SDQ at the mean age of 5 years. The children were assigned to three groups "normal", "borderline" or "abnormal", based on total difficulties score and the borderline and abnormal groups were then combined into a single group. The importance of features was tested with permutation and generalization performance of the models was tested with five fold cross-validation by using stratified subsamples. Accoring to the results the overall performance of the estimator was modest. The best predictors of SDQ emotional symptoms were white matter, amygdala and cerebellar volumes (Area Under the Curve, AUC = 0.74), whereas sex, bronchopulmonary dysplasia and sepsis predicted hyperactivity/inattention symptoms (AUC = 0.75). Combination of socioeconomic risk factors with brain and perinatal factors was found to predict emotional symptoms score (AUC = 0.76), and socioeconomic risk factors correlated with the conduct and peer problems. The limitations of the study included the small sample size and lack of data for term children. [70].

Tate et al. 2020 compared performance of logistic regression, random forest, SVM, neural network and XGBoost in predicting mental health problemns in mid-adolescence at the age of 15 years. Their dataset included 7,638 Swedish twins. The input data included 474 features collected from the register data and parental questionnaire studies and the outcome was SDQ questionnaire data completed by parents. The SDQ total score was transformed into a binary variable by using threshold validated for the Swedish population. According to this threshold approximately 10 % of the adolescences were above the clinical cut-off for mental health problems. Features which were redundant, had no variance or included $\geq$ 50% missing values were excluded from the input data. Missing values were

imputed by using Multivariate Imputation by Chained Equations (MICE) R package. Features with low variance were aggregated when possible. According to their results there were no significant differences in the performance of different models. For random forest and SVM the AUC was 0.74, neural network 0.71, logistic regression 0.70 and XGBoost 0.69, with overlapping 95 % confidence intervals. All the other models except neural networks did not benefit from additional data preprocessing, such as scaling or hyper-parameter tuning. For neural networks these were required. The most important features explaining the outcomes were mental health symptoms (opposite defiant, impulsivity, inattention, executive function and emotional symptoms) reported by the parents, neighbourhood deprivation, peer problems, parity and gestational age at birth and separation anxiety. [103]

Barbosa et al. used ElasticNet penalized regression to examine association of 27 cord blood cytokines with the SDQ outcomes 5-6 year old children (N=869) participating in the French national mother-child cohort (EDEN). Mother's age, BMI, depression, smoking, intake of caffeine and alcohol during pregnancy, symptoms of prenatal anxiety, multiparity, parental education, child's gestational age, mode of delivery, birth weight and sex were included as covariates. Missing values were imputed with MICE package. For modelling the children were categorised into two classes by using 85th upper percentile as a threshold for total difficulties and 15th lower percentile as a threshold for the prosocial subscale. According to the results levels of 12 cytokines associated with at least one of the SDQ subscales. Positive association was found with emotional symptoms (CXCL10, IL10, IL12/IL23p40), peer problems (CCL11, IL17A) and conduct problems (CCL11). In addition, negative association was found with emotional symptoms (IL7, IL15, TNFB), conduct

problems (CCL4, IL6), peer problems (CCL26, IL15) and abnormal prosocial behaviour (CCL26, IL7, IL15, TNFA). [9]. In another similar study of 786 mother-child dyads the found association of IL6 with increased risk and CCL3 and IL16 with decreased risk of prosocial behavior. IL7 was associated with increased risk and IL8, IL10, and IL17A with decreased risk of emotional problems. IL15 was associated with increased and CXCL10 with decreased peer problems. TNFA was associated with conduct problems and CCL2 with hyperactivity/inattention. [10]

2.7 Gaps in knowledge

Emerging evidence suggests that disturbances in the maternal immune, metabolic and hormonal balance during pregnancy can influence neurodevelopment of the offspring. However, large scale studies addressing these questions are still lacking. In particular studies with comprehensive coverage of potentially important features and applying machine learning to predict the outcomes seem not to exist. Addressing these questions has been challenged by the lack of suitable datasets. Furthermore, multiple factors are likely to influence neurodevelopmental health, however, at present it remains unclear which attributes are pivotal and robust in predicting the outcomes. Therefore, traditional statistical approaches, which require preselection of the covariates and are prone to overfitting might not be the optimal approach for addressing these questions.

3 Objectives and aims

3.1 Objectives and aims

The research aim of this study is to determine whether serum biomarkers can be used to predict the behavioural and socio-emotional problems of the children participating in the FinnBrain Birth Cohort Study, and to identify factors in the early life environment, which may co-influence the outcomes. To accomplish this aim, machine learning is used as it enables control of comprehensive collection of potentially important confounding factors and identification of important features collectively contributing to the outcomes.

The data mining aim of the project is to generate approved models which can determine whether serum biomarkers predict the BITSEA or SDQ outcomes and can be used to identify the most important features explaining the outcomes. The biomarkers available for the study include a panel of metabolic, hormonal and immune activation markers measured from maternal serum at gestational week 24 and from children's own serum at the five year follow-up. The outcomes have been assessed at three different ages with two different questionnaire studies. In more detail the study consists of six specific objectives which are to determine whether

A) maternal serum biomarkers or B) children's own serum biomarkers:

1. predict outcomes of the children at the two year follow-up assessed by using Brief Infant-Toddler Social and Emotional Assessment (BITSEA).
2. predict the outcomes of the children at the four year follow-up assessed by using Strengths and Difficulties Questionnaire (SDQ).
3. predict the outcomes of the children at five year follow-up also assessed with Strengths and Difficulties Questionnaire (SDQ).

To achieve the aims and objectives, guidelines from the CRISP-DM framework were followed as applicable.

3.2 Expected results

The results are expected to reveal how accurately we can predict the behavioural and socio-emotional problems of the children based on the given serum biomarkers. Also other factors which may explain the outcomes can be uncovered. Identification of such features would be valuable in facilitating discovery of protective and risk factors for neurodevelopmental outcomes. Furthermore, such factors can also enable identification of individuals at risk and development of effective strategies for early therapeutic intervention. Early intervention would be important for improving quality of life of the affected children and their families as well as for relieving global burden caused by neurodevelopmental and psychiatric disorders. The key factors which may influence the success include project understanding, data quality, quantity and representativeness, data preprocessing approach as well

as model selection and optimisation.

4 Data understanding

4.1 Collection of the datasets

The FinnBrain Birth Cohort Study dataset consists of epidemiological data from nearly 4000 families [58]. At the time of this study the data had been collected from thousands of features starting from gestational week of 14 until the children were in the average age of 5 years. The main input data of interest was the maternal gestational biomarkers and the entire dataset was selected based on the availability of these measurements. The secondary input data of interest was biomarkers measured from the children's serum at the five year follow-up. The datasets were obtained from the FinnBrain Birth Cohort Study in .sav format. The dataset was received in two parts, which were otherwise similar, but the other one contained immune activation marks for the mothers and another for the children. In addition, metadata describing the variables was obtained. The .sav files were converted to .csv files by using R version 4.2.1 and packages readr_2.1.3 and haven_2.5.1:

```
library(haven)
library(readr)
```

```
savToCsv <- read_sav("filename.sav")
write_csv(savToCsv, file="filename.csv")
```

4.2 Description of the datasets

The csv files were uploaded to Jupyter Notebook inital examination. Before cleaning the dataset included gestational biomarkers contained a total of 6,051 feature columns and 1642 rows. The rows represented the mothers and columns the features corresponding to all the variables that were available in the FinnBrain database for these individuals. Of the features 11 were serum biomarkers. For children data was available for 604 individuals (rows) and 11 biomarkers.

4.2.1 Maternal biomarker input data

The maternal biomarkers (Table 4.1) had been measured from the serum, which was collected at the gestational week 24. No fasting was required before sample collection. Before measurements the serum had been stored as frozen. The marker panel included Cholestrol, High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Triglycerides, C-Reactive Protein (CRP), Apolipoportein A (ApoA), ApoB, Glucose, Insulin, Thyroid Stimulating Hormone (TSH) and Thyroid Hormone (FT4). The distribution of biomarker measurements did not follow Gaussian distribution (Shapiro-Wilk test p-value below 0.05 for all). The levels of many of these markers, such as CRP, cholestrol and triglyserides, can be different from the standard reference ranges during pregnancy.

ApoA1 and Ins measurements were missing from one mother and TSH measure-

Table 4.1: Descriptive statistics of gestational serum biomarkers.

Marker	N	Mean	Std	Min	25 %	50 %	75 %	Max
ApoA1 g/l	1641	2.11	0.28	1.15	1.92	2.09	2.27	3.72
ApoB g/l	1642	1.15	0.27	0.53	0.96	1.12	1.31	2.25
Chol mmol/l	1642	6.39	1.07	3.6	5.62	6.3	7.1	11.6
CRP hs mg/l	1642	4.26	3.31	-0.23	1.64	3.27	6.06	12.79
Gluc HK2 mmol/l	1642	4.5	0.85	1.2	4	4.4	4.9	9.9
HDL mmol/l	1642	2.02	0.4	0.74	1.74	2.01	2.28	3.56
LDL mmol/l	1642	3.52	0.92	1.27	2.88	3.43	4.06	7.41
Trigly mmol/l	1642	2	0.7	0.69	1.52	1.87	2.35	5.64
TSH mU/l	1636	1.64	6.07	0.01	1.03	1.34	1.82	245
FT4 pmol/l	1642	10.29	1.46	3.49	9.34	10.1	11	23
Ins mU/ml	1641	32.89	32.85	0.2	10.83	21.39	42.88	244.6

N: number of individuals, Std: standard deviation.

ments from five mothers. In addition, for 24 mothers had error code for below and 175 above detection limit for CRP measurements.

4.2.2 Children's biomarker input data

Children's serum biomarkers (Table 4.2) had been measured at the five year follow-up of the FinnBrain Birth Cohort Study. The markers included Cholestrol, HDL, LDL, Triglycerides, CRP, ApoA, ApoB, Glucose, Insulin, which were the same in mothers and in addition Leptin and Adiponectin, which were not measured from the mothers. TSH and FT4 measured from mothers were not measured from children. The data was not normally distributed based on Shapiro-Wilk test (p-value below 0.05 for all) or visual examination of the data. Some of the biomarkers had strong correlation with each other. Such biomarkers included ApoA1 and HDL (Spearman's rho 0.91), ApoB and LDL (rho 0.9) and LDL and cholestrol (rho 0.9).

For CRP measurements 326 children had an error code for below and 12 above

Table 4.2: Descriptive statistics of children's biomarkers.

Marker	N	Mean	Std	Min	25 %	50 %	75 %	Max
ApoA1 g/l	600	1.49	0.19	0.88	1.36	1.47	1.59	2.18
ApoB g/l	600	0.69	0.13	0.33	0.59	0.67	0.77	1.15
Chol mmol/l	600	4.28	0.68	2.6	3.8	4.3	4.7	7
CRP hs mg/l	600	0.59	1.53	-0.34	0	0.2	0.56	12.6
Gluc HK2 mmol/l	600	4.73	0.76	2.8	4.3	4.7	5.1	11.1
HDL mmol/l	600	1.56	0.3	0.61	1.35	1.53	1.75	2.66
LDL mmol/l	600	2.32	0.53	0.83	1.96	2.28	2.59	4.29
Trigly mmol/l	600	0.92	0.43	0.25	0.62	0.83	1.12	3.78
Ins mU/l	600	16.36	14.7	0.32	7.47	12.62	21.12	168.1
Leptin ng/ml	600	2.38	2.47	0.1	1	1.64	2.83	24.66
Adiponectin ug/ml	600	7.92	2.61	1.99	5.95	7.91	9.54	17.83

N: number of individuals, Std: standard deviation.

detection limit.

4.2.3 BITSEA outcome data

The BITSEA questionnaire study was performed at the two year follow-up and data was available for 941 - 945 children (Table 4.3). The subscales that were used to evaluate BITSEA outcomes included COMPETENCE and PROBLEM scores. Also scores for Autism Spectrum Disorder (ASD) and Pervasive Developmental Disorder (ASD-PDD) subscale were available, however, were not included, as COMPETENCE score has been previously found to detect symptoms of autism spectrum disorders equally well and had a strong negative correlation with ASD-PDD scores (Spearman's rho -0.8, p-value 4.94e-207) also in this study [61].

Table 4.3: Descriptive statistics of BITSEA subscales.

Subscale	N (Tot/M/F)	Mean (Tot/M/F)	Std (Tot/M/F)	Min (Tot/M/F)	Max (Tot/M/F)
COMPETENCE	942/486/456	18.08/17.7/18.49	2.45/2.62/2.2	6.6/6.6/11.0	22.0/22.0/22.0
PROBLEM	939/484/455	7.62/7.92/7.3	4.32/4.16/4.46	0.0/0.0/0.0	34.0/25.0/34.0

Tot: total, F: female, M: male, N: number of individuals, Std: standard deviation.

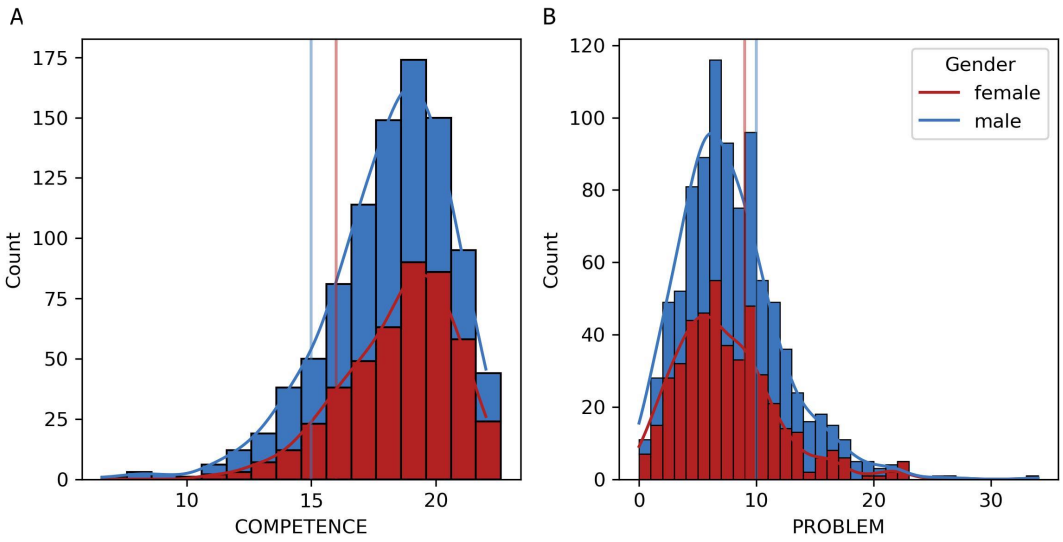

Figure 4.1: Distribution of A) COMPETENCE and) PROBLEM scores from the BITSEA questionnaire study. For COMPETENCE score the vertical lines indicate the sex specific cutpoints of the lowest (dashed line) 15th percentiles. For the PROBLEM score the cutpoints are the 75th percentiles.

According to the visual examination (Figure 4.1) and Shapiro-Wilk test (0.94, p-value 0 for both), the COMPETENCE and PROBLEM scores did not follow normal distribution. Distributions for both subscales were skewed.

High scores in the BITSEA problem total scale or low scores in competence scale may indicate problems. For the PROBLEM subscale a score within the 25th upper percentile suggests a "possible problem" whereas for COMPETENCE subscale a score within the lower 15th percentile suggests "possible deficit/delay range" [3]. In the FinnBrain Birth Cohort Study the lower percentile cutpoint for COMPETENCE score is 15 for boys and 16 for girls, whereas the upper 75th percentile cutoff for PROBLEM score is 10 for boys and 9 for girls (Figure 4.1 and Table 4.4). The number of children in the lowest 15th percentile for competence score

was rather low (85 boys and 84 girls), which may not be sufficient for the training and testing of the models, if the effect sizes are not strong. Due to the inherent nature of the scoring system the outcome data is imbalanced, if binarised to categorical variable by using single cutpoint.

Table 4.4: Cutpoints for BITSEA COMPETENCE and PROBLEM scores and number of children within the lower 15th and in the upper 75th percentiles, respectively.

Subscale	Sex	Cutoff percentile	Cutpoint score	N at risk	Total N
COMPETENCE	boys	15th	15	83	484
COMPETENCE	girls	15th	16	84	455
PROBLEM	boys	75th	10	139	484
PROBLEM	girls	75th	9	158	455

N: number of individuals.

4.2.4 Strengths and Difficulties outcome data

The SDQ questionnaire study had been performed at both four and five year follow-ups. Data from the four year follow-up data was available for 747 - 748 children, and from the five year follow-up for 949 children, (Table 4.5). Similarly to the BITSEA data, the SDQ scores were not normally distributed based on visual examination (Figure 4.2) and Shapiro-Wilk test (at four year follow-up 0.95, p-value 0 and at five year follow-up 0.96, p-value 0).

Table 4.5: Descriptive statistics of SDQ total difficulties scores.

Follow-up	N (Tot/M/F)	Mean (Tot/M/F)	Std (Tot/M/F)	Min (Tot/M/F)	Max (Tot/M/F)
4 years	747/396/351	9.11/9.76/8.38	4.71/4.83/4.46	0.0/1.0/0.0	28.0/26.0/28.0
5 years	949/499/450	8.71/9.3/8.06	4.85/4.91/4.71	0.0/0.0/0.0	27.0/27.0/25.0

Tot: total, F: female, M: male, N: number of individuals, Std: standard deviation.

The total PROBLEM score can be used as a continuous variable or categorical

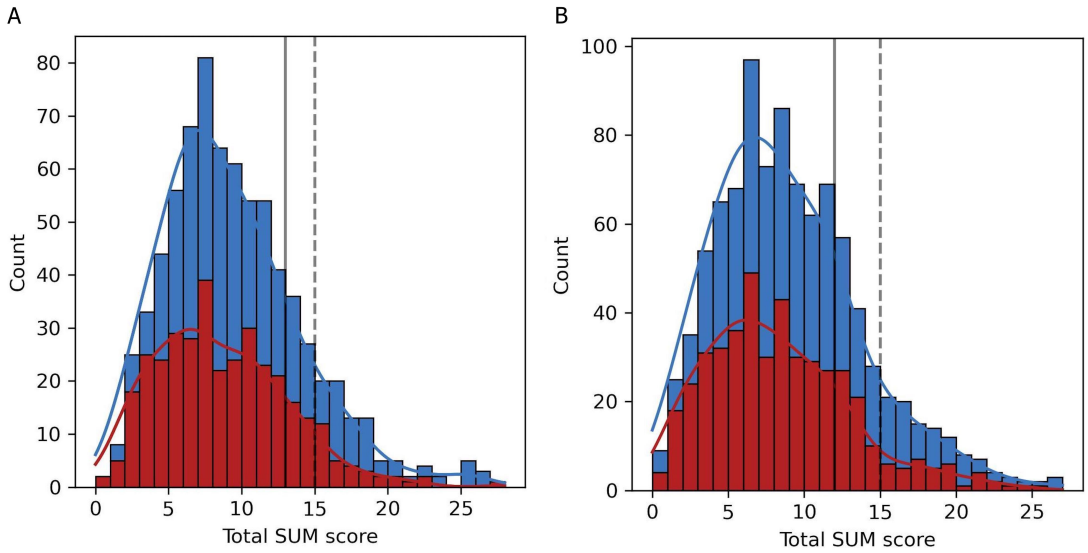

Figure 4.2: Distribution of total difficulties (SUM) scores from the strengths and difficulties questionnaire study at A) four and B) five year follow-ups. The cut-points for upper 80th (solid) borderline and 90th (dashed) percentiles used to identify children at increased risk of problems are indicated with vertical lines.

variable, which can be ranked into either three or four classes. The original cut-points have been established based on UK population. According to the 3-band categorisation (normal, borderline, abnormal), 20 % of the children are expected have score above cutpoint for borderline and 10 % above cutpoint for abonormal group. The 4-band categorisation (close to average, slightly raised/lowered, high-/low, very high/low), has further divided abnormal group into to 5 % categories for better distinction. In the original UK version, for 4-17 years old, the total difficulties (SUM) score rated by parents ranges from 0-13 for normal or close to average groups. The range for borderline and slightly raised/lowered groups is 14-16 and children with score 17-40 belong to the abnormal group, or according to the 4-band system scores 17-19 belong to the high/low group and 20-40 to the very high/low group. [96].

In the FinnBrain Birth Cohort Study the 90th percentile cutpoint is 14 at both timepoints and the 80th percentile cutpoint is 13 at 4 years follow-up and, 12 at 5 years follow-up (Table 4.6). This suggests that the cutpoints are lower in the Finnish population. Similar findings have been previously reported by Borg et al. 2014. According to their study the cutpoint for the top 20 % was 9.5 and top 10 % was 12. With these cutpoints total score had good performance (AUC 79 for 80th percentile cut off, AUC 89 for the 90th percentile cut off) in predicting the severity of symptoms. [17]. The sample numbers for the borderline and abnormal groups were rather small. Similarly to the BITSEA, due to the inherent nature of the scoring system, also SDQ outcome data is also skewed if used as categorical outcome.

Table 4.6: SDQ total difficulties scores and number of children below and above indicated cutpoints.

Follow-up	80th	90th	Normal N	Borderline N	Abnormal N
4 years	13	15	591	63	93
5 years	12	15	712	126	111

N: number of individuals, Std: standard deviation.

4.2.5 Other input features

The dataset contained also numerous other input attributes. These included both numeric and categorical features. Categorical features included both ordinal and nominal features, such as sex of the child. Numeric features seemed to include continuous, such as height or age, and discrete, such as number of fetuses. The features included background of the families, such as birth date, sex, height, weight, postal code, family characteristics, and information about health and medications

of the child and parents collected through questionnaires and from registries. In addition, the other attributes included questionnaire data collected at different time points (Table 4.7).

Table 4.7: Questionnaire studies included in the input dataset and minimum number of responses per question before cleaning.

Abbreviation	Description	gwks 14	gwks 24	gwks 34	birth	3 mo	6 mo	12 mo	14 mo	24 mo	30 mo	4 yrs	5 yrs
QOL	WHOQOL-8 WHO Quality of Life	1546	-	1522	-	-	-	-	-	-	-	-	948
ASS	Anxiety symptom scale	-	-	-	-	-	-	-	-	-	-	-	-
SCL	Anxiety SCL-90 Symptom Checklist-90, 10-item	1544	1575	1515	-	1348	1175	-	-	886	-	734	945
AIS	Athens insomnia scale	1404	1463	1460	1144	864	-	-	-	-	864	-	-
BNSQ	Basic Nordic Sleep	-	-	-	-	-	-	-	-	-	-	-	-
CDRISK	Connor-Davidson Resilience Scale, 10-item	1555	-	-	-	-	-	-	-	-	-	-	952
EPDS	Edinburgh Postnatal Depression Scale	1438	1577	1517	-	1348	1178	1029	-	894	-	735	949
HASSLE	Daily worried and delights	1485	1505	1465	-	-	-	-	-	-	-	-	911
MDAS	Modified Dental Anxiety Scale	1550	-	1521	-	1339	-	-	-	887	-	-	-
MFAS	Maternal–Fetal Attachment Scale	1531	1544	1477	-	-	-	-	-	-	-	-	-
PAI	Use of intoxicants	0	-	0	-	0	-	-	-	-	-	-	-
PRAQ	Pregnancy-Related Anxiety	335	1576	1509	-	-	-	-	-	-	-	-	-
TADS	Trauma and Distress Scale	0	-	-	-	-	-	-	-	-	-	-	-
ECR	Experiences in Close Relationships	-	1572	-	-	-	-	-	-	-	-	-	-
PBI	Parental Bonding Instrument	-	1548	-	-	-	-	-	-	-	-	-	-
SOC	Sense of coherence	-	1585	-	-	1332	-	1021	-	903	-	-	-
PPRFQ	Prenatal parental reflective functioning	-	-	1392	-	-	-	-	-	-	-	-	-
OCHIP	OHIP-14 Oral Health Impact Profile-14	-	-	-	-	-	-	-	-	-	-	-	-
RDAS	Revised dyadic adjustment scale	-	-	1485	-	-	1144	1004	-	857	-	687	862
ETAP	Life events and medicines	-	-	0	-	-	-	-	-	-	-	-	-
VAIKUTE	Impression and experience of the baby	-	-	-	-	1357	-	-	-	-	-	-	-
PBQ	Postpartum Bonding questionnaire	-	-	-	-	1341	-	-	-	-	-	-	-
PSI	Parental Stress Inventory	-	-	-	-	1350	-	1030	-	-	-	730	-
SPSQ	Swedish Parenthood Stress	-	-	-	-	1245	-	939	-	-	-	-	-
BISQ	Brief infant sleep questionnaire	-	-	-	-	-	105	46	-	-	-	-	-
IBQ	IBQ-R Baby's temperament	-	-	-	-	-	400	370	-	-	-	-	-
PRFQ	PoRFQ/To-PRFQ parental reflective func.	-	-	1392	-	-	999	-	-	790	-	-	-
TAS	TAS-20 Toronto Alexithymia Scale	-	-	-	-	-	1171	-	-	-	-	-	937
RUTIIN	Baby's routines	-	-	-	-	-	780	-	-	-	-	-	-
PHOIT	Daycare	-	-	-	-	-	-	40	-	36	-	-	-
ATQ	Adult temperament questionnaire	-	-	-	-	-	-	840	-	-	-	-	-
MC	MCDI MacArthur Communicative Develop.	-	-	-	-	106	-	-	215	-	106	-	-
D	MCDI MacArthur Communicative Develop.	0	1567	1485	188	911	8	40	848	36	911	11	87
MEDIA	Media present in child's daily life	-	-	-	-	-	-	-	-	923	-	11	-
BITSE	Brief Infant Toddler Social Emotional	-	-	-	-	-	-	-	-	16	-	-	-
ECBQ	Early Childhood Behavior	-	-	-	-	-	-	-	-	723	-	-	-
WHO	The Sleep Disturbance Scale for Children	-	-	-	-	-	-	-	-	898	-	752	960
SDSC	The Sleep Disturbance Scale for Children	-	-	-	-	-	-	-	-	904	-	747	954
NEURO	Neurological diseases in relatives	-	-	-	-	-	-	-	-	117	-	-	-
SDQ	Strengths and Difficulties	-	-	-	-	-	-	-	-	-	-	288	385
LIIK	Physical activity of the child	-	-	-	-	-	-	-	-	-	-	272	-
CBQ	Children's Behavior	-	-	-	-	-	-	-	-	723	-	722	658
EA	Interaction between parent and child	1524	1554	1392	1193	1245	999	939	-	790	-	736	-
SUU	Oral health	-	-	-	-	-	-	-	-	-	-	9	-
KOUL	Education	-	-	-	-	-	-	-	-	-	-	-	27
TULOT	Livelihood	-	-	-	-	-	-	-	-	-	-	-	127
ELAMANTAP	Life events	-	-	-	-	-	-	-	-	-	-	-	153
PUHE	Speech and language development	-	-	-	-	-	-	-	-	-	-	-	0
HAMMAS	Child's fear of dental care	-	-	-	-	-	-	-	-	-	-	-	946

4.3 Data quality

Quality of the data was examined. Missing values were detected (Figure 4.3). Due to high number of features assessing whether the values were missing completely at

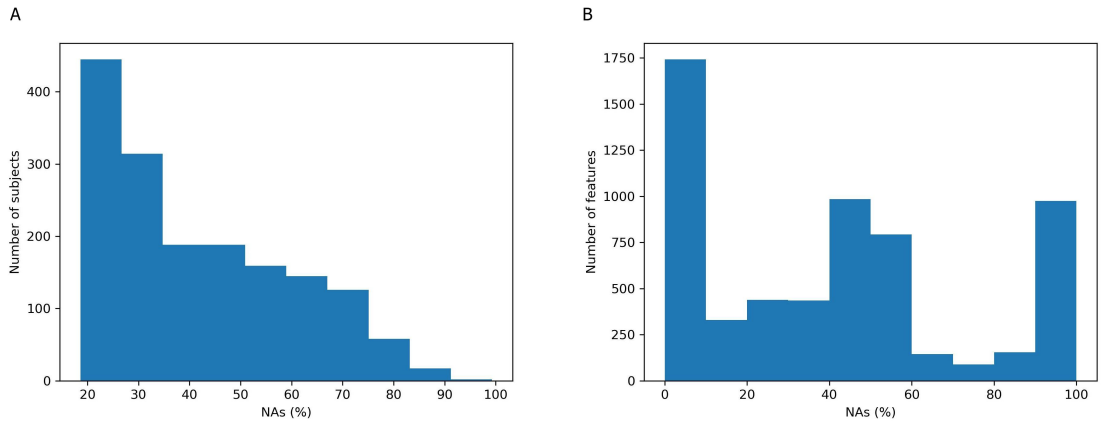

Figure 4.3: Number of A) observations and B) features with different percentages of missing values before cleaning.

random, at random or not at random was found to be a challenge. Nevertheless, there was a possibility that in many of the substudies answers to the sensitive questions, such as those concerning mental health or financial status, were not missing at random and therefore were not ignored. Reason for missing biomarker data was requested from the laboratory providing the measurements. The TSH values were missing for 5 mothers, and APOA1 and Insulin were missing for one mother because the serum sample had run out during the analyses. In addition, as mentioned above, for CRP measurements 24 mothers and 326 children had an error code below, and 175 mothers and 12 children above the detection limit.

Otherwise, the serum biomarkers and output features of interest seemed to be correct, except of one outlier value in maternal TSH measurement. In addition, these attributes were rather clean, except that there were missing values and non-informative features associated with the measurements. Furthermore, the initial dataset contained large number of other non-informative features, such as the date

when questionnaire form has been filled or how the subjects have been divided in substudies. Use of some features was prohibited based on the metadata table from the FinnBrain Birth Cohort Study. Some cells of the dataframe included mixed information and contained extra text or characters. There were also many features which included random and non-systematic answers to open questions. Also redundant features were found, such as data collected from different sources or through repeated questionnaires. In addition, some attributes had been generated from the other features present in the data, such as different types of sum variables or percentages.

As the number of features was so high it was challenging to make definite conclusions of the accuracy of all the attributes based on initial examination. It was also unclear whether there were biases or outliers in the input data. Many of the questionnaire data has categorical variables and in the dataset sum scores had been generated, which are commonly used in statistical modelling and interpreting the outcomes or making diagnosis. In this study the original responses from the questionnaire studies were used instead of sum scores as they may be more informative and also allow assignment of missing values into a dedicated category. This approach may be more informative than imputation.

To conclude the data per se seemed to be relevant for addressing the study questions. The data was not harmonized, contained impurities, redundancies, non-informative features, missing values and was not in a suitable format for modelling. Therefore cleaning and harmonization of the data was performed before further exploration. As for further analyses the data required preparation, other quality aspects, such as correlation, outlier detected and clustering of the date

were examined during or after cleaning.

5 Data preparation

5.1 Selection and cleaning of data

To facilitate data cleaning, selection and preprocessing of the data, an auxillary table was prepared based on the data variable descriptions available from the FinnBrain Birth Cohort study data management team. The auxillary table contained categories (1-12) for each feature data type and expected minimum and maximum values of the feature if available. The categories of the features were:

1. use denied, incomplete
2. sum variable (for PRAQ the recommended sum variables were used as use of individual responses was denied)
3. non-informative
4. duplicate or redundant
5. categorical, string, nominal
6. date
7. numerical (continuous, discrete or ratio not separated, as this would have been too time consuming)

8. categorical, number, nominal
9. year
10. categorical, ordinal, bigger the better, unclear cases classified as 8
11. random free text or no systematic coding
12. categorical, ordinal, bigger the worse, unclear cases classified as 8
13. output categorical variable
14. output recommended numeric sum variable

Different types of errors possibly present in the data were collected and a plan for the data selection, cleaning and preprocessing was prepared by following the guidelines from CRISP-DM and Berthold et al. 2010 [27, 15]. During cleaning prohibited features, non-relevant features, such as FinnBrain's substudy information and redundant features, such as redundant columns of child's sex were excluded from the dataset. Also non-unique features, which had same value in each row were excluded.

Individuals with over 50 % missing values were excluded from the dataset. Empty spaces were removed from the cells, encoding of missing values was harmonized and string features were capitalized. Mistakes and inconsistencies were corrected for features in string format. Extra characters and text were removed from numeric attributes. Representative features or aggregates of features $\geq$ 85 % correlation were selected. The selection was performed so that the feature with the lowest number of missing values was always selected as representative one. After cleaning there were 1425 observations and 1851 features, including outcome related

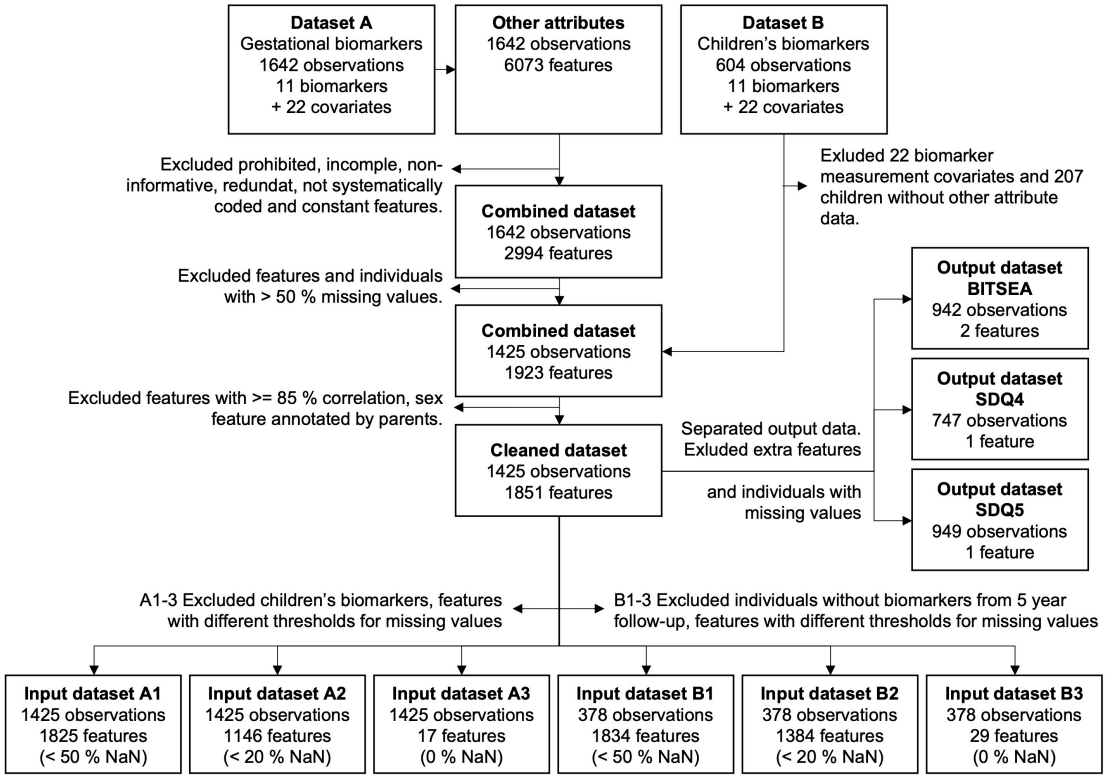

Figure 5.1: Cleaning and management of datasets for modelling. In addition, to removal of features and individuals, the cleaning process involved harmonisation and improvement of accuracy of the data content. See the main text for more details. SDQ = Strengths and Difficulties Questionnaire, BITSEA = Brief Infant Toddler Social Emotional Assessment.

variables, were left in the dataset (Table 5.1, Figure 4).

5.2 Data formatting and construction

No new attributes were produced during preparation of the data. The features which had known expected scale or range of values were examined to detect potential ambiguities. Some features, such as gender, had more categories than were expected. Gender was curated into a feature representing biological sex with two

Table 5.1: Questionnaire studies included in the input dataset and minimum number of responses per question after cleaning.

Abbreviation	Description	gwks 14	gwks 24	gwks 34	birth	3 mo	6 mo	12 mo	14 mo	24 mo	30 mo	4 yrs	5 yrs
QOL	WHOQOL-8 WHO Quality of Life	1359	-	1391	-	-	-	-	-	-	-	-	940
ASS	Anxiety symptom scale	-	-	-	-	-	-	-	-	-	-	-	-
SCL	Anxiety SCL-90 Symptom Checklist-90, 10-item	1358	1395	1385	-	1323	1165	-	-	885	-	-	937
AIS	Athens insomnia scale	1228	1293	1336	1011	-	-	-	-	-	-	-	-
BNSQ	Basic Nordic Sleep	-	-	-	-	-	-	-	-	-	-	-	-
CDRISK	Connor-Davidson Resilience Scale, 10-item	1365	-	-	-	-	-	-	-	-	-	-	944
EPDS	Edinburgh Postnatal Depression Scale	1365	1397	1388	-	1323	1168	1025	-	893	-	-	941
HASSLE	Daily worried and delights	1308	1335	1339	-	-	-	-	-	-	-	-	903
MDAS	Modified Dental Anxiety Scale	1366	-	1392	-	1321	-	-	-	887	-	-	-
MFAS	Maternal–Fetal Attachment Scale	1362	1395	1381	-	-	-	-	-	-	-	-	-
PAI	Use of intoxicants	1327	-	1247	-	810	-	-	-	-	-	-	-
PRAQ	Pregnancy-Related Anxiety	-	1407	1391	-	-	-	-	-	-	-	-	-
TADS	Trauma and Distress Scale	1286	-	-	-	-	-	-	-	-	-	-	-
ECR	Experiences in Close Relationships	-	1401	-	-	-	-	-	-	-	-	-	-
PBI	Parental Bonding Instrument	-	1368	-	-	-	-	-	-	-	-	-	-
SOC	Sense of coherence	-	1402	-	-	-	-	-	-	-	-	-	-
PPRFQ	Prenatal parental reflective functioning	-	-	1384	-	-	-	-	-	-	-	-	-
OCHIP	Oral Health Impact Profile-14	-	-	-	-	-	-	-	-	-	-	-	-
RDAS	Revised dyadic adjustment scale	-	-	1365	-	-	1145	1010	-	860	-	-	859
ETAP	Life events and medicines	-	-	1322	-	-	-	-	-	-	-	-	-
VAIKUTE	Impression and experience of the baby	-	-	-	-	1332	-	-	-	-	-	-	-
PBQ	Postpartum Bonding	-	-	-	-	1318	-	-	-	-	-	-	-
PSI	Parental Stress Inventory	-	-	-	-	1325	-	1026	-	-	-	-	-
SPSQ	Swedish Parenthood Stress	-	-	-	-	1316	-	1016	-	-	-	-	-
BISQ	Brief infant sleep	-	-	-	-	-	1138	966	-	-	-	-	-
IBQ	IBQ-R Baby's temperament	-	-	-	-	-	-	-	-	-	-	-	-
PRFQ	PoRFQ/To-PRFQ parental reflective func.	-	-	1384	-	-	1153	-	-	894	-	-	-
TAS	TAS-20 Toronto Alexithymia Scale	-	-	-	-	-	1161	-	-	-	-	-	930
RUTIIN	Baby's routines	-	-	-	-	-	-	-	-	-	-	-	-
PHOIT	Daycare	-	-	-	-	-	-	929	-	881	-	-	-
ATQ	Adult temperament	-	-	-	-	-	-	837	-	-	-	-	-
MC	MCDI MacArthur Communicative Develop.	-	-	-	-	-	-	-	899	-	-	-	-
D	MCDI MacArthur Communicative Develop.	1286	1397	1365	-	902	1145	1010	899	860	902	747	859
MEDIA	Media present in child's daily life	-	-	-	-	-	-	-	-	922	-	-	-
BITSE	Brief Infant Toddler Social Emotional Ass.	-	-	-	-	-	-	-	-	939	-	-	-
ECBQ	Early Childhood Behavior	-	-	-	-	-	-	-	-	831	-	-	-
WHO	The Sleep Disturbance Scale for Children	-	-	-	-	-	-	-	-	897	-	-	952
SDSC	The Sleep Disturbance Scale for Children	-	-	-	-	-	-	-	-	903	-	-	946
NEURO	Neurological diseases in relatives	-	-	-	-	-	-	-	-	-	-	-	-
SDQ	Strengths and Difficulties	-	-	-	-	-	-	-	-	-	-	747	949
LIIK	Physical activity of the child	-	-	-	-	-	-	-	-	-	-	-	-
CBQ	Children's Behavior	-	-	-	-	-	-	-	-	831	-	-	-
EA	Interaction between parent and child	-	-	-	-	-	-	-	-	939	-	-	-
SUU	Oral health	-	-	-	-	-	-	-	-	-	-	-	-
KOUL	Education	-	-	-	-	-	-	-	-	-	-	-	-
TULOT	Livelihood	-	-	-	-	-	-	-	-	-	-	-	909
ELAMANTAP	Life events	-	-	-	-	-	-	-	-	-	-	-	926
PUHE	Speech and language development	-	-	-	-	-	-	-	-	-	-	-	889
HAMMAS	Child's fear of dental care	-	-	-	-	-	-	-	-	-	-	-	939

categories based on data from healthcare register. For other questionnaires extra unexpected category was left intact as they may be informative and can present an answer in which the participant did not know how to reply. Language of the child was also present in several columns and for some of the individuals the answer varied. The columns were merged into a single feature by selecting the most common answer as the final one.

The CRP values marked with an error code "20" (below detection limit) or "21"

(above detection limit) were left intact as they were close to the observed minimum and maximum values of the feature, respectively. Distributions of all numerical features were visualized to detected potential outliers (Figure 5.2). One exceptionally high value (245 mIU_L) was observed for the feature "Result_TSH_mIU_L" (Figure 5.2a). The value was curated to the second largest value (9.192 mIU_L) observed for the feature (Figure 5.2b). There seemed to be variation also for many other attributes, however, not as exceptional. As it was not clear based on the data what is true variation, rest of the features were left intact.

To increase number of children with SDQ data available, the possibility to combine four and five year old time points was examined. Although there was clear correlation between the scores, it was not very strong (Spearman rho 0.73, p-value $\geq$0.01). Furthermore, merging would have increased the number of individuals only with 5 %. Therefore, the time points were not combined. After cleaning of the data, appropriate formats of the features were ensured and transformation was performed when needed. All ordinal variables were converted to categorical format. The datasets were stored in .pkl file format to preserve the data types of the features.

5.3 Splitting of the data to training and test sets

After cleaning the data was separated into training (65 %) and test (35 %) sets. These thresholds were chosen with the aim to ensure a sufficient number of observations in the test dataset. The output and input features were separated for the preprocessing and modelling. The training dataset was used for model optimisa-

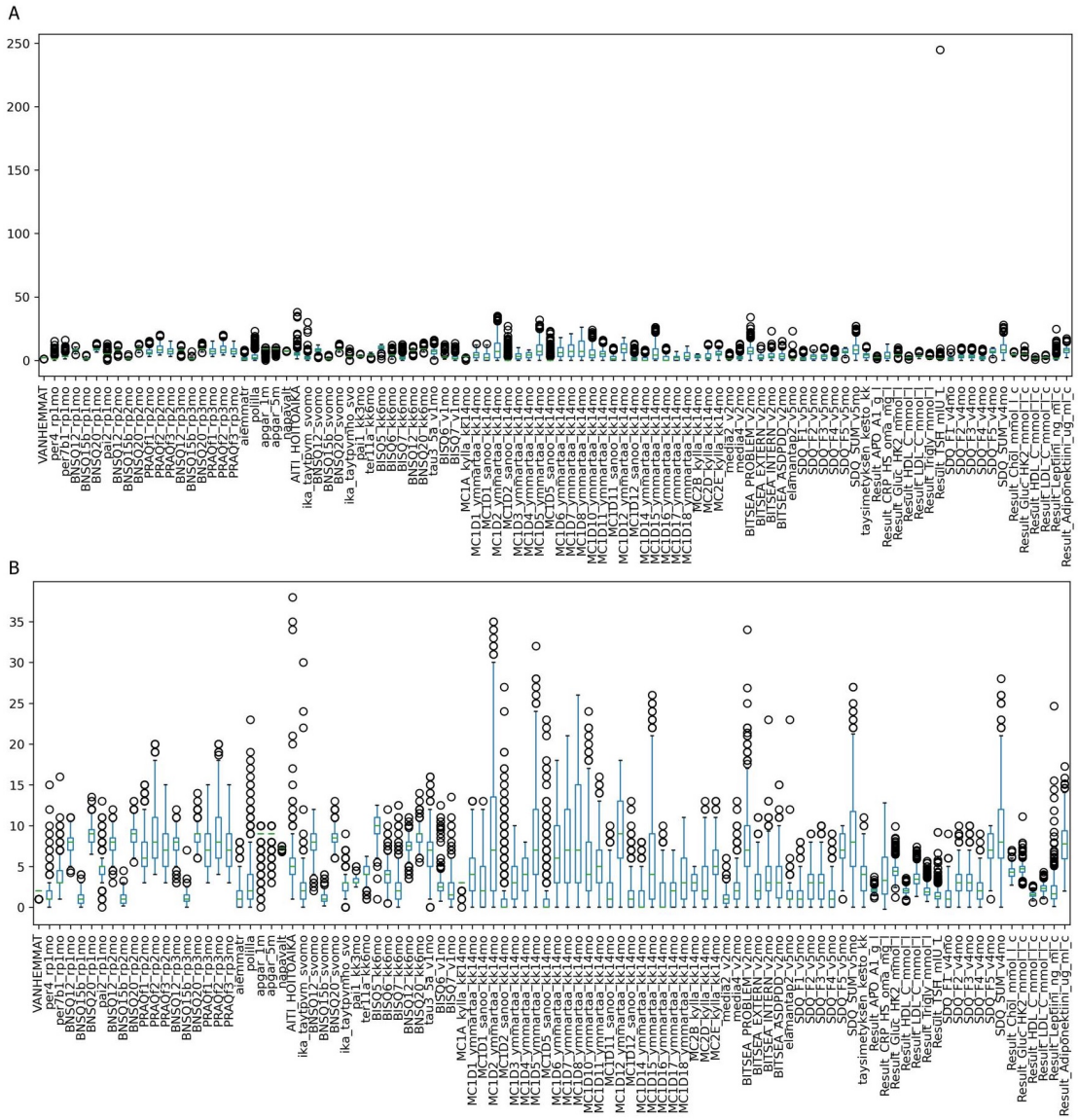

Figure 5.2: An example of value distributions of numeric features before and after curation of an outlier. In the figure is a subset of numeric features with mean values above 5 and less or equal to 10. The box-and-whisker plots in figure A) show the distributions of numeric values before, and figure B) after curation of an outlier in TSH measurements. The box represents the distribution of data from first to third quantiles. Median value is indicated by the horizontal line within the box. The whiskers represent the range of the data within 1.5-times the interquartile range. Observations outside this range (outliers) are displayed as dots.

tion and selection by using five-fold cross-validation. The test dataset was saved for evaluating the generalisation performance of the final models.

5.4 Principal component analysis

Principal component analysis (PCA) was performed on the training input dataset to examine number of components explaining the variation in the data, to visualise data structures, patterns, outliers and to detect possible issues in the data. Two different imputers, inclucing KNN Imputer and Simple Imputer as well as four different scaling methods, including Standard Scaler, Robust Scaler, MinMax Scaler and MaxAbs Scaler from Sckikit-learn modules [84] were used to also get insight into how these different preprocessing methods influence the data structures. Input dataset for the analysis was the training dataset (Input dataset A1 from figure 5.1) allowing 50 % of missing values in features, excluding children's biomarker data and observations that were not present in SDQ5 output dataset. Test dataset was not included in the analysis as the results were used to select the most promising tools for preprocessing. Based on the PCA the distributions for other approaches except combination of Robust Scaler and Simple imputer were very similar with two distinguishable main clusters (Figure 5.3). In combination of Simple Imputer and Robust Scaler some pronounced outliers appeared in the PCA. KNNImputer and RobustScaler were chosen for the further preprocessing. In addition to insensitivity to outliers, Robust Scaler has been previously shown to have good overall performance [1], although for classification tasks, the performance of different algorithms may vary depending on the scaling technique [4].

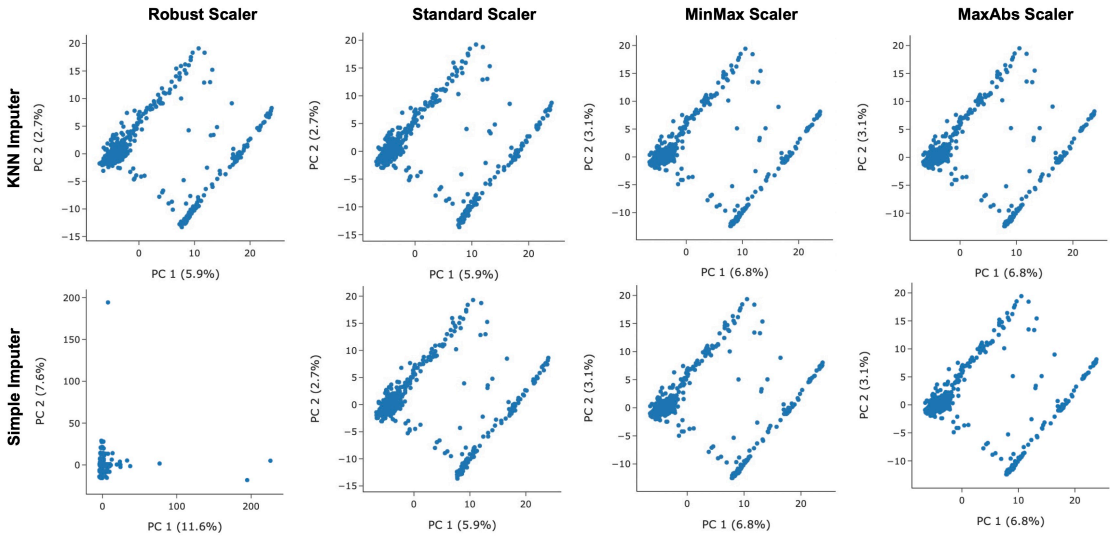

Figure 5.3: Principal Component Analysis of the training dataset. Influence of different scalers and imputers on the training dataset (Input dataset A1 from figure 5.1) structure was examined with Principal Component Analysis (only components 1 and 2 are shown). The input dataset includes features with up to 50 % missing values and excluded children's biomarkers. Only the observations present in output dataset Strengths and Difficulties at five year follow-up (SDQ5) were included in the analysis.

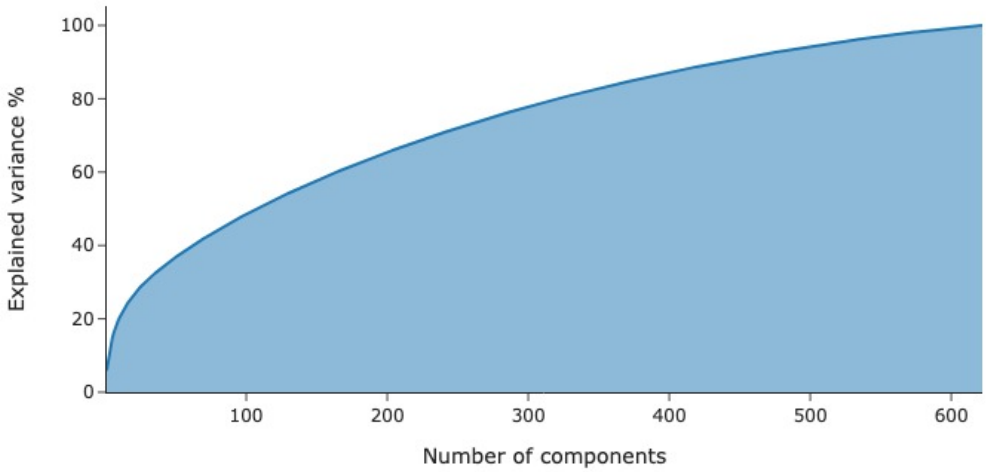

Figure 5.4: Number of features required to explain variance in the input dataset. Principal Component Analysis was performed by using the input dataset, which includes the features with up to 50 % missing values, and excluded children's biomarkers and observations not present in the five year Strengths and Difficulties follow-up output dataset. Cumulative sum of the explained variance with increasing number of features was calculated and is show in the figure.

When using KNN Imputer in combination with Robust Scaler, component 1 (PC1) explained 5.9 % of the total variance in the data and component 2 (PC2) explained 2.7 % of the variance (Figure 5.3). Calculation of cumulative sum of the percentage of explained variance based on PCA analysis revealed that there are no individual features which would explain large proportion of the variance and that most of the features present in the data are needed to explain the variance (Figure 5.4).

5.5 Preprocessing pipeline

A preprocessor function was generated to be implemented later in the modelling pipeline. Preprocessing steps were performed by using the scikit-learn's pipeline

to prevent any leakages between the training and validation data [84]. The preprocessing steps included imputation of numerical missing values by using scikit-learn's KNN Imputer, scaling by using Robust Scaler and encoding of categorical features to numeric format by using OneHotEncoder. A dedicated category was generated for the missing categorical attributes. In addition, categories with less than 20 observations were categorised as infrequent observations.

6 Modelling

The modelling and selection of the machine learning technique was performed as an iterative process. As the aim was to predict neurodevelopmental outcomes of the children by using high-dimensional input data and several output features, a supervised approach was chosen. According to the literature, the scores for both BITSEA and SDQ outcomes can be treated as continuous scores or categories. For categorisation the cutpoints can vary based on gender, population and responder. Although previous literature suggests that the cutpoints for Finnish children differ from the original studies, standardised thresholds at the population level have not been established yet. Therefore, the regression was chosen as the first approach. An alternative approach was to label the children based on cutpoints chosen by the percentiles available in the guidelines, however, the sample numbers above or below the recommended cutpoints were rather low and due to scoring system the outcome data was also imbalanced when categorised. For each outcome only the features collected until corresponding age of assessment were included in the input dataset.

6.1 Selection of evaluation metrics

6.1.1 Evaluation metrics for regressors

The performance of the regressors was primarily assessed with coefficient of determination (R-squared or R^2). R^2 is the default metric used for evaluation of many of the regression models in the scikit-learn libraries. R^2 quantifies the goodness of fit of the model's regression line thereby indicating the the proportion of variance in the outcome feature that is predictable based on the input features. In the mathematical notation of R^2 (1) $\hat{y}$ denotes for the predicted i^{th} value, y^i represents the true i^{th} value, $\bar{y}$ is the mean of observed true values and n is the total number of observations. [33]

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2} \tag{6.1}$$

The benefits of R^2 metric include it's range from values -inf to 1. Values ≤ 0 indicate no correlation between the input and output features whereas value 1 indicates for perfect fit. Therefore the metric is easier to interpret than many other commonly used metrics, such as mean average error (MAE), mean squared error (MSE), square root of mean square error (RMSE) or mean absolute percentage error (MAPE), which can have infinite positive values. When evaluating the models R^2 values $\geq$0.7 were considered to suggest fitting of the model whereas values $\leq$0.3 were be considered as a weak or no fit. In parallel, other metrics, such as MSE (2), RMSE (3) and mean absolute error (MAE) (4) were be used as needed.

[33]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{6.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{6.3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{6.4}$$

For RMSE and MAE value 0 indicate perfect fit whereas positive values which can be +inf indicate worse than perfect fit. Each metric has limitations and advantages. For example RMSE is more sensitive to outliers than MAE [33], whereas RMSE may perform better than MAE when the errors follow gaussian distribution [26]. On the other hand, R^2 has also criticized to be insufficient and sometimes misleading estimator of accuracy. [110]. Therefore, use of several metrics is typically needed.

6.1.2 Evaluation metrics for classifiers

Classification was considered as an alternative approach. Commonly used metrics to evaluate classifier performance include for example, accuracy (4), precision (5), recall (6), F1-score (7), Matthews correlation coefficient (MCC) (8) and AUC [90]. The metrics can be mathematically defined as

$$Accuracy = \frac{TN + TP}{FN + FP + TN + TP} \tag{6.5}$$

$$Precision = \frac{TP}{FP + TP} \tag{6.6}$$

$$Recall = \frac{TP}{FN + TP} \tag{6.7}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6.8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{(FN + TN) \times (FN + TP) \times (FP + TN) \times (FP + TP)} \tag{6.9}$$

where TN denotes for the number of true negative observations, TP for true positives, FN for false negatives and FP for false positives. The scores for precision, recall, accuracy, F1 and Receiver Operating Characteristic (ROC)-AUC range from 0 to 1 corresponding to worst and best fit, respectively. Previous studies on health outcomes have approved models with accuracy or AUC score above 0.65 providing

some guidelines of the expected performance. The MCC values range from 1 to -1 and can be interpreted similarly to the Pearson correlation coefficient (r). Value 1 indicates perfect prediction, whereas -1 indicates perfect inverse prediction and 0 indicates no prediction. [90]

These metrics were collected for all the models, however, as a primary metric for model selection and evaluation, MCC together with F1 and ROC-AUC scores were used. Benefits of these metrics are that they are more tolerant to imbalanced datasets, which is a common issue in classification tasks. The dataset is imbalanced when, for example in binary classification task, there are many more observations for the so called majority or positive class than for the minority or negative class. If the dataset is imbalanced, accuracy tends to give misleadingly over-optimistic estimates, for example by classifying all the observations into the majority class. While F1 and MCC may be more tolerant to imbalances, in certain cases they can also provide misleading results. For example, when the data size of the minority class is small, MCC may not perform well, although in general has been found to have robust performance. F1 again is sensitive to class swapping and by default does not consider true negatives in the scoring. [18, 108, 32, 23]. Therefore, similarly to assessment of regression models, evaluation of model performance based on multiple metrics is a good practise.

6.2 Selection of algorithms

As it was not clear which algorithms would have the best performance with FinnBrain data representative algorithms from different families were tested as

a first step to find the most optimal one. As the primary aim was to test whether maternal and children's own biomarkers markers predict the outcomes in the presence of unknown covariates or confounding factors, and it was important to know which factors potentially explain the outcomes, algorithms, which provide information of the explanatory features were considered as the first approach. Only algorithms which do not make strong assumptions of the normality, multicolinearity or heteroskedasticity of the data were considered. The relationship between the input and output features was not necessarily linear and there could be interactions and hidden relationships between the input features. Therefore, algorithms capable of modelling both linear and non-linear relationships, and capturing hidden relationships were included. The size of the dataset was small or medium, so algorithms requiring large datasets, such as deep learning were considered not to be optimal for this study. As the data was high dimensional, only algorithms enabling penalization were considered.

6.2.1 ElasticNet

ElasticNet regressor (ENET) was selected as it has been found to perform well with high dimensional data and in the presence of multicolinearity. It has previously been succesfully use to predict SDQ outcomes based on cord blood cytokine levels [9]. ElasticNet models linear relationships between input and output features. To minimise risk of overfitting penalization is enabled by using hybrid of L1 and L2 regularisation terms. The ratio of penalty can be adjusted with L1 ratio parameter. L1 ratio 0 is equal to L2 penalty and 1 is equal to L1 penalty. The values between are combination of the penalty terms. Parameter α of ElasticNet regressor can

be used to adjust the weigth of penalty. [99, 118] In the scikit learn the objective function to minimise the training loss by ElasticNet is defined as

$$\begin{aligned} &\text{Objective} = \text{MSE loss} + \alpha \times \ell_1 \text{ ratio} \times \ell_1 + 0.5 \times \alpha \times (1 - \ell_1 \text{ ratio}) \times \ell_2 \\ &\ell_1 = \lambda \sum_{j=1}^{p} |\beta_j| \\ &\ell_2 = \lambda \sum_{j=1}^{p} \beta_j^2 \end{aligned} \tag{6.10}$$

where λ controls the strength of penalty, p denotes for the number of features in the model and β is the value of j^{th} coefficient in the model [84].

6.2.2 Generalized Linear Regressor

Generalized Linear Regressor was selected as it has been shown to perform well with different types of datasets [39], it provides information of the explanatory features and enables modelling of non-linear relationships through transformation of features. Therefore, the outcome feature can follow for example binomial, poisson or gamma distribution. However, also this approach assumes linear relationship between input and output features after transformation. To implement generalised linear models TweedieRegressor (TWR) from Sckikit-learn was be used. TweedieRegressor enables modelling of tweedie distributions, however, by adjusting the power parameter distributions, such as Poisson (power = 1), Compound Poisson Gamma (power = 1,2), Gamma (power = 2) or Inverse Gaussian (power = 3) can be modelled. The mathematical notation for generalised linear models is [84]:

$$\text{Objective} = \min_{w} \frac{1}{2n_{\text{samples}}} \sum_{i} d(y_i, \hat{y}_i) + \frac{\alpha}{2}\ell_2, \tag{6.11}$$

where d denotes for the unit deviance function associated with the distribution of the exponential family and α denotes for the weigth of L2 penalty. [84].

6.2.3 Logistic regression

In case regressors do not perform well, logistic regression (LGR) was considered as an alternative to ElasticNet and TweedieRegressor. Logistic regression has been widely applied in classification tasks with good performance. It has also one of the most popular models in epidemiological studies [12, 79, 103, 106].

Scikit implementation of logistic regression enables automatic feature selection and control of overfitting with L1, L2 or elastic net regularisation terms [84]:

$$\text{Objective} = \min_{w} C \sum_{i=1}^{n} (-y_i \log(\hat{p}(\hat{y}_i = 1|x_i)) - (1 - y_i) \log(1 - \hat{p}(\hat{y}_i = 1|x_i))) + r(w) \tag{6.12}$$

where $\hat{p}(\hat{y}_i = 1|x_i$ is the probability that the i^{th} instance of predicted target sample $\hat{y}$ belongs to the positive class, y is the true observed label, C is the regularisation term and $r(w)$ denotes the regularisation term ℓ_1, ℓ_2 or ElasticNet.

6.2.4 Support Vector Machine

Support Vector Machine was chosen based on the previously reported good performance [39] and popularity in previous epidemiological studies [12, 79, 106, 103,

70]. Support Vector Machines can be applied to solve both regression and classification problems. [35, 98]. In classification approach one or more hyperplanes are generated and fitted in the high- or infinite-dimensional space by maximising the separation between the classes. The margin for separation is maximised by selecting the hyperplane with the greatest distance to the nearest observation (support vector). In regression task the aim is to find the best fit rather than separation of classes. [94, 113, 104]. According to the Scikit learn, the objective function for Support Vector Regressor (SVR) with linear kernel can be formulated as follows:

$$\begin{aligned} &\text{Objective} = \min_{w,b} \frac{1}{2} w^T w + C \times \epsilon\text{-insensitive loss}, \\ &\epsilon\text{-insensitive loss} = \sum_{i=1}^{n} \max(0, |y_i - (w^T \phi(\hat{y}_i) + \beta)| - \varepsilon), \end{aligned} \tag{6.13}$$

where w denotes for weights and b the bias of the intercept of the fitted hyperplane. $\frac{1}{2} w^T w$ is the regularisation term, C is the penalty for the loss function. In the loss function y_i is the i_{th} observed value for the target and $w^T \phi(\hat{y}_i) + \beta$ is the linear function for the predicted i_{th} target value $\hat{y}$ and $\phi(\cdot)$ represents the kernel trick enabling modelling of both linear and non-linear relationships between output and input features in high-dimensional space. Kernel parameters that can be specified include for example 'linear', 'poly', 'rbf' and 'sigmoid'. The ϵ parameter must be non-negative and it denotes for the distance around the predicted values, so called ϵ-tube, where the errors are considered to be non-significant and no penalty is applied for the training loss. [84]. As an alternative to regressor, Support Vector Classifier (SVC), was considered. Previous studies have found it to have good

general performance in classification tasks [25, 40, 117].

6.2.5 Extreme Gradient Boosting algorithm

Extreme Gradient Boosting (XGB) is a state-of-the-art ensemble algorithm, which was introduced in 2016. It demonstrated good performance in several different studies and machine learning competitions. It has previously been used to predict SDQ outcomes in Swedish twin study [103]. In the most common implementation decision trees are used, however, modelling with neural networks is also possible. XGB can solve both regression and classification tasks. It exploits gradient decent to minimise loss function, and L1 and L2 penalisation terms to minimise risk of overfitting. XGB has a cache-aware block structure enabling parallelization. Other benefits of XGB include that it can manage well large datasets, provides feature importances and is easy to interpret. [30, 117, 13, 94]. Therefore, XGBRegressor (XGBR), and if needed XGB Classifier (XGBC), from Scikit-learn were also included for comparison. The disadvantage of the algorithm is the large number of hyperparameters which may need to be optimised. The simplified objective function for XGB can be specified as [114]:

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \hat{y_i}^{(t)}) + \sum_{i=1}^{t} \omega(f_i) \tag{6.14}$$

where $L(\cdot)$ denotes for the loss function for the i^{th} value of true observed target sample y and predicted target sample $\hat{y}$ for ensemble tree t. The total number of samples is denoted by n. $\omega(\cdot)$ denotes for the regularisation term for i^{th} tree

f, which includes the tree structure and leaf scores. The total number of trees is denoted by t. As an alternative to regressor, XGBoost classifier was considered, if needed. [30, 117, 13].

6.2.6 K Nearest Neighbor algorithm

KNN algorithm was also included in the comparison as it has been s common choice in previous epidemiological studies[12]. It is an instance based learning algorithm, which can be applied in both classification and regression tasks. It is capable to solve non-linear problems. The main components of the algorithm are distance metric, neighbor count, weighting scheme and prediction function. For classification KNN generates n-dimensional space where it stores the observations in the training dataset. New observations are classified based on the majority vote of similarity with the nearest neighbors. The number of nearest neighbors to be considered is specified by the parameter k. The similarity is measured with the chosen distance metric, such as Euclidean or Minkowski distance. In regression, the outcome is predicted by calculating an average of the output values for the K number of neighbors. Weight of the nearest neighbors in comparison to distant ones can be adjusted. For optimal performance, the parameters, such as number of neighbors and distance metric need to be optimised. Benefits of KNN include that it is robust when the training data is noisy. Limitations include sensitivity to the data quality. [15, 94].

6.3 Test design and construction of models

The training dataset (65 % of the whole data) was used for optimisation, training and comparison of the algorithms. The outcomes of interest included two subscales of BITSEA questionnaire study, PROBLEM sand COMPETENCE scores from two year follow-up, and total difficulties score (SUM) from the SDQ questionnaire study performed at four and five year follow-ups. The input dataset was analysed with and without childrens' biomarkers, as childrens' biomarkers were available only from a subset of mother-child pairs. For these input dataset three different thresholds (0, 20 and 50 %) were used for allowing missing values in the input features. For biomarkers missing values were allowed only for those individuals who had participated in the laboratory testing. For each outcome, only those features that had been collected up to the time point of assessment were included in the analysis.

The selected algorithms were optimised and trained by using pipeline including preprocessing and gridSearchCV from the Scikit-learn library [84]. In this approach the hyperparameter tuning was performed in the inner loop and model performance and selection was evaluated in the outer validation loop through repetative process by using the default 5-fold cross-validation. The evaluation metrics described in section 6.1 were used to assess model performances. This gridSearchCV approach was chosen as a first step as it was not clear which type of algorithms and hyperparameters would best model the relationships in the data. The approach enabled automated hyperparameter tuning, efficient use of the training dataset and simultaneous comparison of several algorithms in replicated design. The constraints of this approach were that limited amount of information was obtained

of the behaviour of the models during optimisation. Furthermore, due to rather high computational cost, the number of hyperparameters that could be tested was not high. Therefore, as a first approach, coarse search was performed in the parameter space. After identifying the most promising model architecture, more comprehensive optimisation was performed for the chosen algorithm.

6.4 Model optimisation

6.4.1 Performance of regressors

As a first step in model optimisation, the targets were treated as a continuous variables and performances of five different regressors (ENET, KNR, SVR, TWR and XGBR) in predicting the targets were compared. According to the results none of the regerssors included in the analyses predicted the outcomes. The mean validation R^2 scores varied from -0.28 to 0.20 indicating poor performance of the models (Table 6.1). The best possible score for R^2 would be 1.0, which indicates perfect fit or prediction, whereas negative values indicate that the model can be arbitrarily worse. Score of 0 indicates that the model predicts always the expected average of output features despite of the input attributes. Transformation of the target features to more normal like distribution did not improve the performance (data not shown).

As a conclusion, the regression models did not learn to predict the outcome from the data with the given parameters. The validation scores were at so low level that sufficient further improvement through more thorough optimisation seemed unlikely.

Table 6.1: Performance of regressors in predicting the BITSEA COMPETENCE and PROBLEM scores and SDQ total difficulties scores (SUM) (R^2 validation scores).

Subscale	age	CBM	NaN %	ENET mean (std)	KNR mean (std)	SVR mean (std)	TWR mean (std)	XGBR mean (std)
COMPETENCE	2 yrs	no	0	0.00 (0.01)	0.03 (0.02)	0.00 (0.02)	0.02 (0.03)	-0.18 (0.11)
COMPETENCE	2 yrs	no	20	0.02 (0.03)	0.00 (0.03)	0.03 (0.03)	0.04 (0.03)	-0.18 (0.11)
COMPETENCE	2 yrs	no	50	0.13 (0.03)	0.04 (0.03)	0.12 (0.03)	0.17 (0.04)	0.00 (0.14)
PROBLEM	2 yrs	no	0	0.00 (0.01)	-0.01 (0.01)	-0.02 (0.03)	0.00 (0.02)	-0.21 (0.14)
PROBLEM	2 yrs	no	20	0.10 (0.03)	-0.03 (0.04)	0.06 (0.02)	0.11 (0.02)	0.01 (0.06)
PROBLEM	2 yrs	no	50	0.20 (0.04)	0.00 (0.07)	0.11 (0.07)	0.19 (0.03)	0.14 (0.07)
SUM	4 yrs	no	0	-0.02 (0.04)	0.00 (0.05)	-0.04 (0.05)	-0.01 (0.05)	-0.23 (0.07)
SUM	4 yrs	no	20	0.10 (0.01)	-0.02 (0.05)	-0.03 (0.08)	0.12 (0.04)	0.01 (0.09)
SUM	4 yrs	no	50	0.13 (0.04)	0.01 (0.05)	0.12 (0.04)	0.16 (0.04)	0.00 (0.05)
SUM	5 yrs	no	0	0.01 (0.01)	0.00 (0.02)	-0.01 (0.02)	0.02 (0.02)	-0.14 (0.11)
SUM	5 yrs	no	20	0.14 (0.04)	-0.01 (0.03)	0.08 (0.03)	0.14 (0.04)	-0.02 (0.04)
SUM	5 yrs	no	50	0.18 (0.05)	-0.01 (0.05)	0.14 (0.04)	0.19 (0.04)	0.07 (0.06)
SUM	5 yrs	yes	0	-0.02 (0.02)	-0.03 (0.03)	-0.04 (0.04)	-0.02 (0.02)	-0.28 (0.15)
SUM	5 yrs	yes	20	0.15 (0.16)	0.00 (0.06)	0.15 (0.18)	0.14 (0.13)	0.09 (0.09)
SUM	5 yrs	yes	50	0.15 (0.18)	0.01 (0.06)	0.17 (0.17)	0.15 (0.14)	0.14 (0.24)

CBM: includes children's biomarkers, NaN%: percentage of missing values allowed for features, std: standard deviation.

6.4.2 Performance of classifiers

Due to the poor performance of regressors, the modelling strategy was revised and classification was used as an alternative approach. For this purpose, the scores of the outcome features, BITSEA COMPETENCE and PROBLEM scores, and SDQ total difficulties SUM scores, were converted to classes. The outcome features were converted to binary variables by using single cutpoints, which were calculated from the data based on the percentile thresholds provided in the BITSEA and SDQ study manuals [22, 96].

For BITSEA COMPETENCE score, the cutpoint was the sex specific lower 15th percentile [22, 3]. In the entire dataset, including holdout dataset, there were 83 boys and 84 girls below the cutpoint. For BITSEA PROBLEM score the cutpoint was in 75th percentile including 139 boys and 158 girls in the entire dataset (Table 4.4). For SDQ data, the number of individuals belonging to the abnormal group

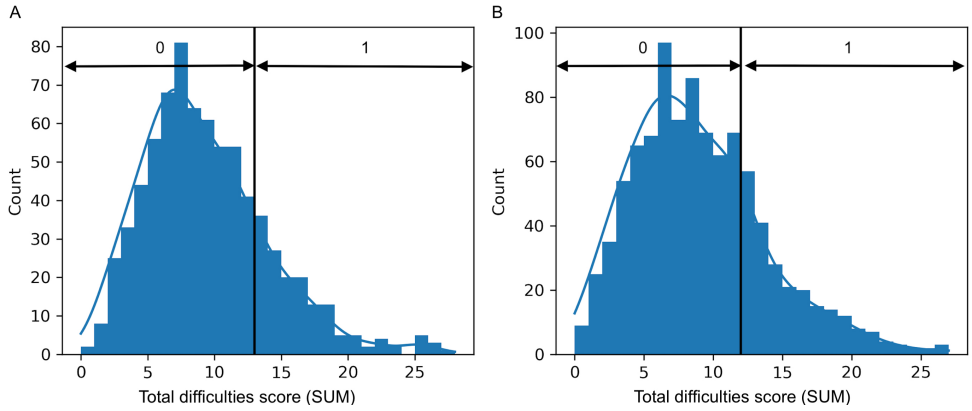

Figure 6.1: Binarisation of Strength and Difficulties (SDQ) total difficulties scores by using single threshold. The total difficulties scores (SUM) from A) four and B) five year follow-ups were binarised by using single cutpoints calculated based on percentile thresholds provided in the SDQ manual. The continuous scores were converted to binary values by using 80 percentile threshold as a cutoff so that scores below the threshold were converted to 0 and scores above the threshold were converted to 1. According to the SDQ manual the scores within the upper 20th percentile are considered as borderline and those within upper 90th percentile are considered abnormal.

(in the 90th percentile) was low. The total number of instances was N=93 at four year follow-up and N=111 at five year follow-up, including both the training and holdout test datasets. Therefore, the 80th percentile, including both borderline and abnormal cases, was chosen as the cutpoint (Table 4.6). Similar approach has also been described previously in the literature. [70]. By using the cutpoints calculated from the percentiles, the outcome scores were converted to binary values so that scores below the outcome specific thresholds were converted to 0 and scores above the thresholds were converted to 1, except for BITSEA COMPETENCE score the labels were opposite (see figure 6.1 for an example). See chapter 2.4.1 for more details how to interpret BITSEA scores, and chapter 2.4.2 for SDQ scores.

After binarisation of the target features, performance of four different classifiers,

KNC, LGR, SVC and XGBC in predicting the outcomes was compared. According to the results, the performance for most of the datasets was not sufficient. The most promising results were achieved with the XGBC in predicting for total difficulties scores with input datasets allowing missing values. In the best model 50 % of missing values were allowed and children's biomarkers were included in the input features. The mean MCC validation score of this model was 0.35, however, the standard deviation was high (0.16) and performance based on other evaluation metrics was modest (F1 mean validation score 0.43, std 0.15 and ROC-AUC mean validation score 0.63, std 0.07). The results for four and five year total difficulties outcomes (SUM) excluding children's biomarkers and allowing missing values were also promising. Performance of all the other models was poor with mean MCC scores ranging from 0 to 0.18 (Table 6.2).

As performance of XGB Classifier seemed the most promising, optimisation was continued by focusing on this algorithm.

6.4.3 Performance of XGB Classifier with minority class weights

To further improve the performance of the models, approach was revised again. Due to the inherent nature of BITSEA and SDQ scoring methods, which aim to identify the minority group of children at-risk or with potential problems, the distributions of target scores were skewed (Figures 4.1, 4.2, 6.1). Furthermore, as described in the literature review, the thresholds for classifying the children into different groups are not clear-cut and may vary in different populations and settings. Machine learning is typically based on the assumptions that the goal of the

Table 6.2: Performance of classifiers in predicting the BITSEA COMPETENCE and PROBLEM scores and SDQ total difficulties scores (SUM) (MCC validation scores).

Subscale	Age	CBM	NaN %	KNC mean (std)	LGR mean (std)	SVC mean (std)	XGBC mean (std)
COMPETENCE	2 yrs	no	0	0.00 (0.00)	-0.01 (0.02)	0.00 (0.00)	0.02 (0.08)
COMPETENCE	2 yrs	no	20	0.00 (0.00)	0.04 (0.05)	0.00 (0.00)	0.05 (0.10)
COMPETENCE	2 yrs	no	50	0.00 (0.00)	0.03 (0.09)	0.00 (0.00)	0.03 (0.09)
PROBLEM	2 yrs	no	0	0.03 (0.12)	0.11 (0.05)	0.19 (0.08)	0.07 (0.04)
PROBLEM	2 yrs	no	20	0.03 (0.05)	0.03 (0.07)	0.00 (0.00)	0.07 (0.08)
PROBLEM	2 yrs	no	50	0.03 (0.05)	0.09 (0.08)	0.00 (0.00)	0.10 (0.08)
SUM	4 yrs	no	0	0.00 (0.00)	0.06 (0.13)	0.07 (0.17)	0.03 (0.06)
SUM	4 yrs	no	20	0.13 (0.05)	0.16 (0.12)	0.00 (0.00)	0.28 (0.07)
SUM	4 yrs	no	50	0.13 (0.09)	0.29 (0.09)	0.00 (0.00)	0.29 (0.11)
SUM	5 yrs	no	0	0.01 (0.11)	0.09 (0.05)	0.16 (0.04)	0.07 (0.05)
SUM	5 yrs	no	20	0.03 (0.04)	0.18 (0.02)	0.00 (0.00)	0.24 (0.1)
SUM	5 yrs	no	50	0.05 (0.04)	0.27 (0.05)	0.00 (0.00)	0.3 (0.1)
SUM	5 yrs	yes	0	0.01 (0.13)	-0.03 (0.08)	0.08 (0.21)	0.18 (0.08)
SUM	5 yrs	yes	20	0.05 (0.1)	0.23 (0.17)	0.00 (0.00)	0.26 (0.07)
SUM	5 yrs	yes	50	0.04 (0.13)	0.19 (0.11)	0.00 (0.00)	0.35 (0.16)

CBM: includes children's biomarkers, NaN%: percentage of missing values allowed for features, std: standard deviation.

process is to maximise the accuracy, and that distribution of the unseen test data will be the same as it is for the training data. Therefore, in the presence of class imbalance, meaning that classes of target the feature are unequally distributed, performance of algorithms can be poor. Severity of the problem is influenced by the degree of imbalance, complexity of the data, size of the training dataset and type of the algorithm used for modelling. [88, 62].

Several strategies have been developed to mitigate the issues caused by class imbalance. Traditional approaches include re-sampling, either downsampling or oversampling, which involve artificial balancing of the target variable. In downsampling, also known as downsizing, majority observations are excluded, whereas in upsampling, also known as oversampling, minority observations are replicated.

Resampling can be performed on random instances. However, the limitations of the random approach are that the undersampling can lead to loss of important information, whereas oversampling can increase the risk of overfitting. Alternatively, also hybrid of approach of undersampling and oversampling can be used. In addition, numerous other extensions and techniques have been developed [88, 62, 109]. For example, the resampling approach can also be directed, involving elimination or generation of new observations based on informed choices or adjusting the decision threshold based on specific criteria. Further alternative approaches include comparison of probability distributions of the model, for example by using all the possible thresholds. Also adjustment of the target threshold to correct one has been proposed. In addition, specific algorithms and their modifications can be used to manage the imbalance, such as one-class learning or weighted outputs [29, 88, 62].

To gain better understanding of the degree of imbalance in this study, the imbalance ratios were calculated for all the outputs as previously described [62]:

$$Imbalance\ Ratio\ (IR) = \frac{Total\ number\ of\ majority\ observations}{Total\ number\ of\ minority\ observations} \tag{6.15}$$

According to the results, the COMPETENCE outcome had IR=5 and PROBLEM had IR=2. SDQ total difficulties outcome at four year follow-up had IR=6 and at five year follow-up IR=7 for the abnormal class. When borderline and abnormal classes were combined the IR decreased to 3 and 2, respectively. If the IR ratio is high, it will be a challenge for the machine learning model to distinguish the

minority class from noise [109]. For the outcome features in this study, the imbalance seemed not to be severe. However, the distributions were skewed and it was not clear whether this hampered the modelling.

One of the techniques to control for imbalance is class weighting. XGB Classifier has a specific hyperparameter, 'scale_pos_weight' for this purpose [114]. The hyperparameter adjusts the weights of the entire positive class, in this case the minority class. The recommended values for the weights are given by the IR formula (6.15). To test whether controlling of the imbalance would improve the performance, this hyperparameter was included in the gridSearchCV with search space ranging from 1 to 8.

According to the mean MCC validation scores, performance of the models predicting SDQ total difficulties and allowing missing values in the input features seemed again the most promising approach (Table 6.3). However, no clear difference or improvement was observed in the performance when compared to the models without class weights. All the models predicting BITSEA outcomes had again poor performance.

6.4.4 Performance of XGB Classifier with hybrid resampling

Minority class weighting did not have clear impact on the model performance. Therefore, as an alternative approach, resampling was tested. For this purpose a hybrid approach was implemented as a part of the preprocessing and grid-

Table 6.3: Performance of XGB Classifier in predicting the outcome scores when using positive class weights (MCC validation scores).

Outcome	Age	CBM	NaN %	F1 mean (std)	MCC mean (std)	ROC-AUC mean (std)
COMPETENCE	2 yrs	no	0	0.1 (0.06)	0.06 (0.09)	0.51 (0.02)
COMPETENCE	2 yrs	no	20	0.07 (0.07)	0.1 (0.14)	0.51 (0.02)
COMPETENCE	2 yrs	no	50	0.17 (0.13)	0.08 (0.14)	0.53 (0.05)
PROBLEM	2 yrs	no	0	0.43 (0.06)	0.14 (0.1)	0.57 (0.05)
PROBLEM	2 yrs	no	20	0.44 (0.03)	0.12 (0.03)	0.56 (0.02)
PROBLEM	2 yrs	no	50	0.31 (0.02)	0.14 (0.03)	0.56 (0.01)
SUM	4 yrs	no	0	0.36 (0.05)	0.07 (0.1)	0.54 (0.06)
SUM	4 yrs	no	20	0.41 (0.08)	0.25 (0.1)	0.62 (0.05)
SUM	4 yrs	no	50	0.46 (0.07)	0.29 (0.09)	0.65 (0.05)
SUM	5 yrs	no	0	0.39 (0.05)	0.13 (0.07)	0.57 (0.04)
SUM	5 yrs	no	20	0.38 (0.08)	0.23 (0.09)	0.6 (0.04)
SUM	5 yrs	no	50	0.45 (0.05)	0.28 (0.07)	0.63 (0.03)
SUM	5 yrs	yes	0	0.39 (0.09)	0.18 (0.16)	0.58 (0.07)
SUM	5 yrs	yes	20	0.47 (0.16)	0.32 (0.2)	0.64 (0.09)
SUM	5 yrs	yes	50	0.49 (0.12)	0.3 (0.17)	0.65 (0.08)

CBM: includes children's biomarkers, NaN%: percentage of missing values allowed for features, std: standard deviation.

SearchCV pipeline by using combination of libraries available from scikit-learn [84] and imbalanced-learn [67]. As a first step in the resampling, the observations in the majority class, residing near the decision boundaries, were cleaned by using Edited Nearest Neighbours method. As a second step, the observations belonging to the minority class were oversampled by using Synthetic Minority Over-sampling Technique (SMOTE), a commonly used method for oversampling. SMOTE creates synthetic examples for each minority class observation by using nearest neighbors approach. [28, 111, 11].

According to all the evaluation metrics, the datasets predicting SDQ total difficulties outcomes (SUM) and allowing missing values in the input datasets were again the most promising ones. The best performance was now gained with the dataset allowing 20 % of missing values and including children biomarkers. How-

ever, overall the performance of the models remained at modest or poor level, and considering the standard deviations, was not clearly improved from the previous approach. Datasets not allowing missing values were excluded from further analysis as it seemed unlikely that those models can be further improved.

Table 6.4: Performance of XGB Classifier after undersampling with Edited Nearest Neighbours followed by oversampling with SMOTE method (validation scores).

Outcome	Age	CBM	NaN %	F1 mean (std)	MCC mean (std)	ROC-AUC mean (std)
COMPETENCE	2 yrs	no	0	0.25 (0.07)	0.06 (0.1)	0.53 (0.06)
COMPETENCE	2 yrs	no	20	0.14 (0.04)	0.02 (0.04)	0.51 (0.02)
COMPETENCE	2 yrs	no	50	0.17 (0.12)	0.03 (0.13)	0.51 (0.06)
PROBLEM	2 yrs	no	0	0.45 (0.03)	0.09 (0.05)	0.54 (0.03)
PROBLEM	2 yrs	no	20	0.47 (0.03)	0.08 (0.07)	0.54 (0.03)
PROBLEM	2 yrs	no	50	0.46 (0.03)	0.06 (0.08)	0.53 (0.04)
SUM	4 yrs	no	0	0.34 (0.04)	0.08 (0.06)	0.55 (0.04)
SUM	4 yrs	no	20	0.42 (0.04)	0.26 (0.06)	0.62 (0.03)
SUM	4 yrs	no	50	0.43 (0.09)	0.27 (0.12)	0.63 (0.06)
SUM	5 yrs	no	0	0.4 (0.02)	0.12 (0.05)	0.57 (0.03)
SUM	5 yrs	no	20	0.45 (0.01)	0.26 (0.02)	0.63 (0.01)
SUM	5 yrs	no	50	0.45 (0.07)	0.28 (0.09)	0.63 (0.05)
SUM	5 yrs	yes	0	0.38 (0.14)	0.1 (0.15)	0.56 (0.08)
SUM	5 yrs	yes	20	0.55 (0.18)	0.37 (0.25)	0.7 (0.13)
SUM	5 yrs	yes	50	0.49 (0.08)	0.28 (0.12)	0.65 (0.06)

CBM: includes children's biomarkers, NaN%: percentage of missing values allowed for features, std: standard deviation.

6.4.5 Performance of XGB Classifier with sample weights

Another approach to manage imbalance is to use sample weighs. In this approach each training instance of the target feature is assigned with a weight based on its importance. In the presence of class imbalance, performance of XGBC has been previously described to improve by sample weighting [115]. Therefore, this approach was tested next. For this purpose, each target observation in the training dataset was assigned with a weight based on their importance. Each instance

belonging to the minority class (class 1) were assigned with weight 1. The instances belonging to the majority class (class 0) were given a weight based on their distance from the cutpoint value before binarisation. The weights for majority class were calculated as follows:

$$\begin{aligned} d_i &= |y_i - c|^2 \\ w_i &= \frac{d_i}{\max\{d_1, d_2, \ldots, d_n\}} \end{aligned} \tag{6.16}$$

where d denotes for the squared absolute distance of i^{th} observation in the sequence of target variables y from the the outcome specific cutpoint c for binarisation, and w denotes for corresponding weight for each instance scaled between 0 and 1. Also weights based on scaled absolute distances and log2 transformed distances were tried, however, the best training-validation performance was obtained with the exponential approach (data not shown).

When using sample weights, performance of the models predicting SDQ outcome and allowing missing values were considered to reach acceptable levels (mean MCC validation score ≥ 0.3), although the overall performances were still modest. The mean MCC validation scores ranged from 0.32 to 0.46, F1 scores from 0.53 to 0.71, and ROC-AUC scores from 0.65 to 0.73 (Table 6.5). The models predicting BITSEA outcomes had again poor performances.

Based on these results SDQ SUM scores from 4 year and 5 year follow-ups were chosen for continuation. BITSEA outcomes were not included in the further analyses as it seemed unlikely that performance could be improved.

Table 6.5: Performance of XGB Classifier with sample weights (validation scores).

Outcome	CBM	Age	NaN %	F1 mean (std)	MCC mean (std)	ROC-AUC mean (std)
COMPETENCE	2 yrs	no	20	0.39 (0.06)	0.07 (0.05)	0.53 (0.02)
COMPETENCE	2 yrs	no	50	0.6 (0.06)	0.11 (0.1)	0.55 (0.05)
PROBLEM	2 yrs	no	20	0.54 (0.04)	0.16 (0.04)	0.58 (0.02)
PROBLEM	2 yrs	no	50	0.62 (0.03)	0.14 (0.06)	0.57 (0.03)
SUM	4 yrs	no	20	0.53 (0.05)	0.32 (0.07)	0.65 (0.03)
SUM	4 yrs	no	50	0.57 (0.06)	0.35 (0.09)	0.66 (0.04)
SUM	5 yrs	no	20	0.63 (0.04)	0.34 (0.05)	0.67 (0.03)
SUM	5 yrs	no	50	0.58 (0.05)	0.35 (0.05)	0.67 (0.03)
SUM	5 yrs	yes	20	0.71 (0.06)	0.46 (0.12)	0.73 (0.06)
SUM	5 yrs	yes	50	0.7 (0.12)	0.46 (0.22)	0.73 (0.11)

CBM: includes children's biomarkers, NaN%: percentage of missing values allowed for features, std: standard deviation.

6.4.6 Performance of XGB Classifier after fine-tuning

As the models predicting SDQ outcomes by using XGBC and sample weights in coarse hyperparameter search seemed the most promising, they were selected for fine-tuning in more comprehensive hyperparameter space (Table 6.6). The optimisation was performed systematically with five-fold GridSearchCV approach by optimising two hyperparameters at the time. The process was repeated, until no clear improvement in the validation scores was gained. MCC was used as the primary metrics to evaluate performance of the models. Alongside with MCC score F1 score and ROC-AUC scores were monitored. During optimisation, the models with the best MCC scores were chosen.

According to the results, the performance of all models was good or moderate. The mean MCC validation scores ranged from 0.37 to 0.59, F1 scores from 0.75 to 0.86 and ROC-AUC scores from 0.69 to 0.75 (Table 6.7).

Table 6.6: Hyperparameter search space used in fine-tuning of XGBoost Classifier.

Parameter	Description	Range (steps)
subsample	subsample ratio of instances	0.5-1.0 (0.05)
colsample_bytree	subsample ratio of tree	0.5-1.0 (0.05)
max_depth	max depth of three	3-9 (1)
min_child_weight	min sum of instance weight in child node	1-20 (1)
n_estimators	number of trees	100-500 (10)
gamma	min split loss	1-20 (1)
reg_alpha	L1 regularisation	1-20 (1)
reg_lambda	L2 regularisation	1-20 (1)
learning_rate	feature weights step size	0.05-0.30 (0.05)

Table 6.7: Performance of XGB Classifier in predicting SDQ total difficulties outcomes after fine-tuning (mean validation scores).

Age	CBM	NaN %	F1 mean (std)	MCC mean (std)	ROC-AUC mean (std)
4 yrs	no	20	0.75 (0.03)	0.42 (0.05)	0.7 (0.03)
4 yrs	no	50	0.76 (0.02)	0.46 (0.05)	0.73 (0.02)
5 yrs	no	20	0.76 (0.07)	0.37 (0.14)	0.69 (0.07)
5 yrs	no	50	0.81 (0.04)	0.42 (0.11)	0.71 (0.05)
5 yrs	yes	20	0.85 (0.11)	0.58 (0.21)	0.75 (0.09)
5 yrs	yes	50	0.86 (0.08)	0.59 (0.15)	0.75 (0.05)

CBM: includes children's biomarkers, NaN%: percentage of missing values allowed for features, std: standard deviation.

Performance of models allowing either 20 % or 50 % of missing values in the input features was similar. Shapiro-Wilk test was used to estimate normal distribution of the MCC validation scores, although there were scores available only from five folds of cross-validation for each dataset. According to the test statistics, the scores for all the other datasets were normally distributed, except for five year dataset allowing up to 50 % missing values in input features and including children's biomarkers (Shapiro-Wilk p-value ≤ 0.05). Performance of models was then compared with non-parametric Mann-Whitney U exact test and parametric independent two-tailed T-test. According to the results there were no significant differences (p-values > 0.05) in the performance of the models (Table 6.8). Based

on these findings, the input data allowing % 20 missing values was chosen for continuation.

Table 6.8: Statistical comparison of MCC validation scores of models including A) 20 % vs B) 50 % of missing values in input features.

CBM	Age	A Shapiro-Wilk p	B Shapiro-Wilk p	Mann-Whitney p	T-test p
no	4 yrs	0.41	0.44	0.42	0.29
no	5 yrs	0.16	0.53	0.55	0.54
no	5 yrs	0.01	0.04	0.84	0.93

CBM: Children's biomarkers included, p: p-value.

6.5 Generalisation performances

The generalisation performances of optimised models were then examined in the holdout test dataset without sample weights. The performance scores were lower than in the training-validation dataset. MCC scores ranged from 0.21 to 0.28 (validation scores 0.37-0.59), F1 scores from 0.63 to 0.65 (validation scores 0.75-0.86) and ROC-AUC scores from 0.63 to 0.65 (0.69-0.75) (Table 6.9). Based on the ROC-AUC scores the performance was considered to be moderate [52]. The possible explanation for the drop of the scores from those observed in training-validation phase could be overfitting. Alternatively, the hold-out test dataset was not representative or large enough as the number of children in the minority groups was rather low.

6.6 Feature importances

The main aim of this study was to determine whether the serum biomarkers predict the BITSEA and SDQ outcomes, and to identify features that may co-influence

Table 6.9: Generalisation performances of the final models in predicting total difficulties outcome from the Strengths and Difficulties questionnaire study.

CBM	Age	NaN %	Observations (N)	Features (N)	F1	MCC	ROC-AUC
no	4 yrs	20	226	1146	0.65	0.21	0.63
no	5 yrs	20	335	1146	0.65	0.28	0.66
yes	5 yrs	20	335	1146	0.63	0.26	0.65

CBM = children's biomarkers, NaN % = percentage of missing values allowed for features, N = number.

the outcomes. For this purpose, the features explaining the model were identified by using two different methods, permutation importance [19] and SHapley Additive exPlanations (SHAP) method [72, 73]. Permutation is a robust method, which reveals the features affecting model's performance by randomly shuffling and removing them from the model one at the time. SHAP was chosen as it has been demonstrated to provide consistent and accurate local explanations by revealing how features interact and influence the outcome at the level of individual predictions. Computation of Shapley values for the entire dataset also helps in understanding models global behaviour and enables identification of features which collectively contribute to the outcome of interest. Furthermore, SHAP method can also reveal directionality how features present in the input data influence the outcome.

6.6.1 Most important features explaining total difficulties outcome at four year follow-up

SHAP method revealed in total of 32 attributes (Shapley value ≥ 0.01) influencing the SDQ total difficulties outcome at four year follow-up. The only serum biomarker among these features was maternal gestational Thyroid stimulating hor-

mone level, TSH mIU/L (mean importance 0.04, std 0.006), although the influence on the outcome seemed to be rather weak (Figure 6.2). Closer examination of the Shapley values revealed that lower maternal TSH levels measured during pregnancy predicted higher total difficulties scores and vice versa.

Several other features also predicted the total difficulties. Factors such as mother not being puzzled about her body at six months follow-up (TAS7), experience that baby can be easily consolidated (PBQ25) and positive attitude towards life (SPSQ28) at three months follow-up predicted lower total difficulties score. Minutes being awake at leisure (BNSQ2b) and symptoms of anxiety (PRAQsum10) during pregnancy, and number of marriages (per4 rp1) predicted total difficulties of the children (Figure 6.2). Description of all 32 features and their influence in total difficulties score at four year follow-up is provided in Appendix A (Table A.1). Permutation test did not find any serum biomarkers among the important features. These results suggest that none of the serum biomarkers alone have strong influence on the four year outcome and that TSH influences the outcome in interaction with the other features.

6.6.2 Most important features explaining total difficulties outcome at five year follow-up

A total for 101 features with mean Shapley value of ≥ 0.01 were found to influence the five year SDQ total difficulties outcome when children's biomarkers were not included in the input dataset. Maternal gestational TSH level was now the most important feature predicting the total difficulties with mean Shapley value of 0.34 (std 0.10). Similarly to the four year's follow-up lower gestational TSH levels

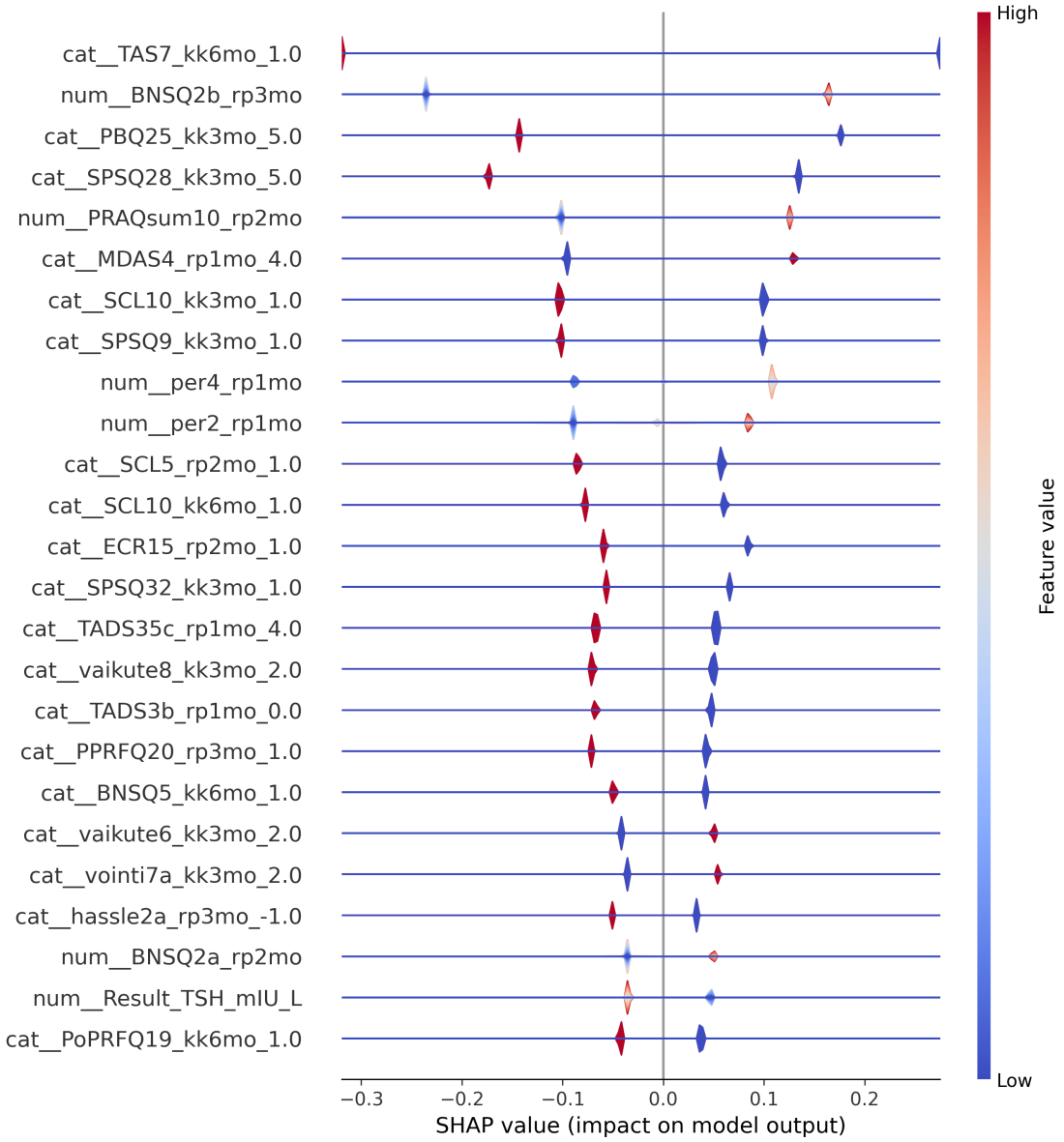

Figure 6.2: Shapley values for the 25 most important features predicting SDQ total difficulties outcome at the four year follow-up. See Appendix A (Table A.1) or more details.

predicted increased risk of total difficulties and vice versa (Figure 6.3).

In addition to TSH, also several other serum biomarkers were found among the factors explaining the five year outcome (Figure 6.4). Of these LDL (mean 0.06, std 0.02), APOA1 (mean 0.05, std 0.03), Trigly (mean 0.03, std 0.04) and FT4 (mean 0.02, std 0.02) were associated with increased risk of total difficulties when gestational levels were low and vice versa. In contrast, gestational Glucose HK2 (mean 0.05, std 0.03) and Insulin (mean 0.03, std 0.05) levels seemed to have opposite pattern, such that high Glucose HK2 levels predicted increased risk of difficulties and lower levels had opposite, but weaker influence. Lower levels of Insulin predicted decreased risk of difficulties, however, higher levels seemed not to have as strong influence on the outcome. LDL levels had very strong correlation with APOB (Sperman's rho 0.90, FDR = 0) and Cholestrol levels (Spearman's rho 0.90, FDR = 0), indicating that these features can have redundant influence on the model. LDL was found as an important feature also by permutation test (mean importance 0.009 +/- 0.004). However, the influence of all these other biomarkers on the outcome was weaker than that of TSH.

In the Finnish population, range 0.07-2.50 mU/l has been suggested as a normal reference interval for TSH levels [76]. In the entire dataset 153 women had a value above this range and only 7 below the range. Of the mothers 101 had thyroxin mediation at gestational week 12, 113 at week 34 and 83 at five year follow-up. Medication, however, did not correlate with TSH levels or SDQ total difficulties score at five year follow-up. Thyroxin use during pregnancy had a weak or modest correlation with FT4 levels. At gwk 14 Point-Bisearial correlation efficient was 0.31 (p-value 1.25e-37) and at week 34 0.26 (p-value 5.17e-27).

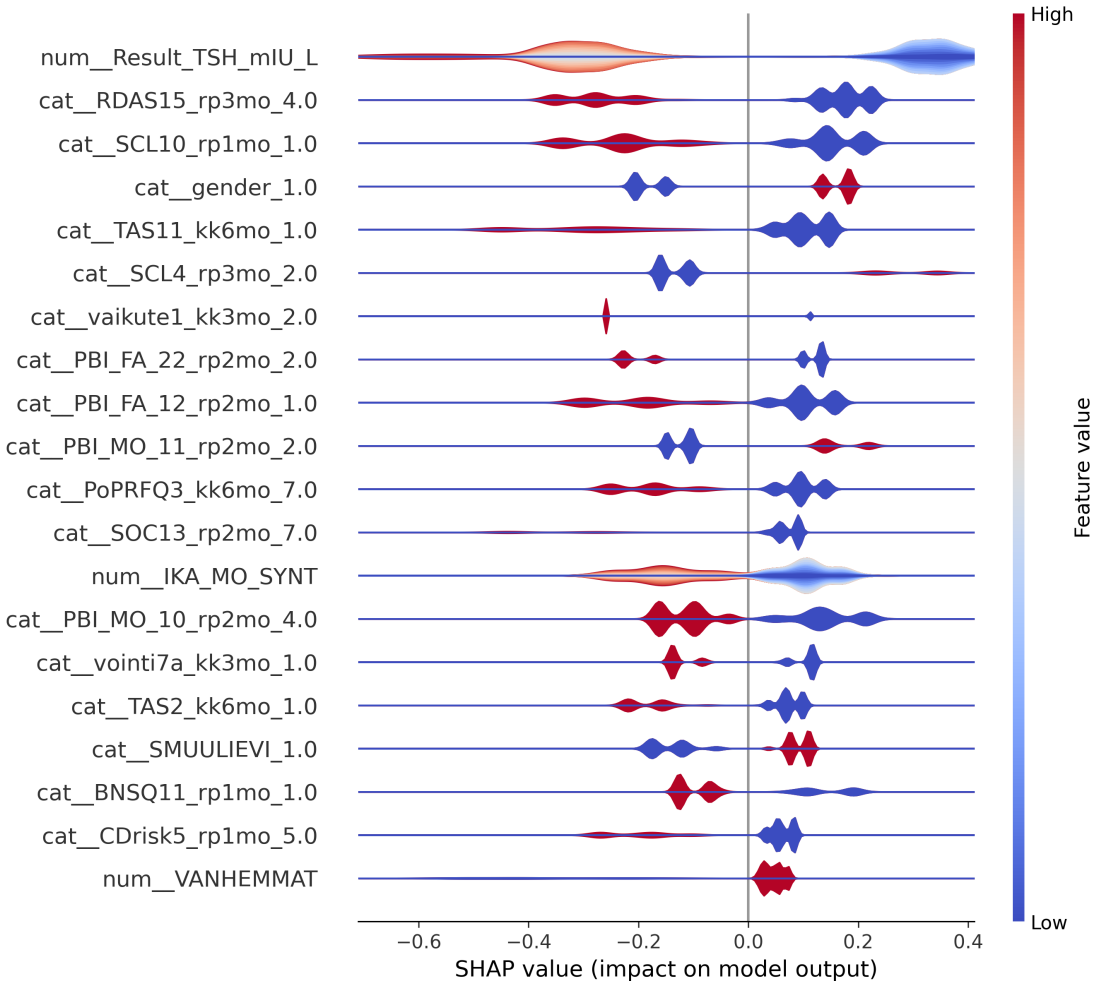

Figure 6.3: Shapley values for the 20 most important features predicting total difficulties outcome at the five year follow-up. See Appendix A (Table A.2) for more details.

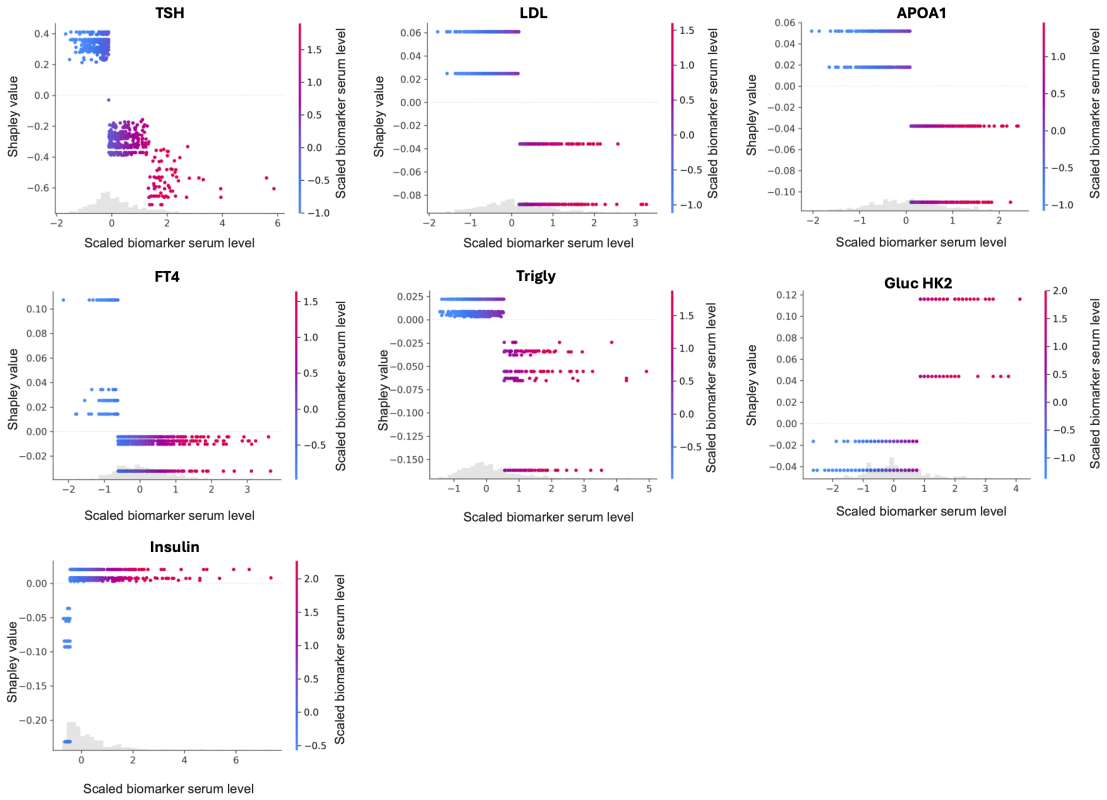

Figure 6.4: Influence of serum biomarkers on Strengths and Difficulties total difficulties outcome at five year follow-up. The Shapley values in the y-axis indicate the strength of feature's influence on the outcome and the scaled serum levels of the biomarker are shown in the x-axis and as coloring.

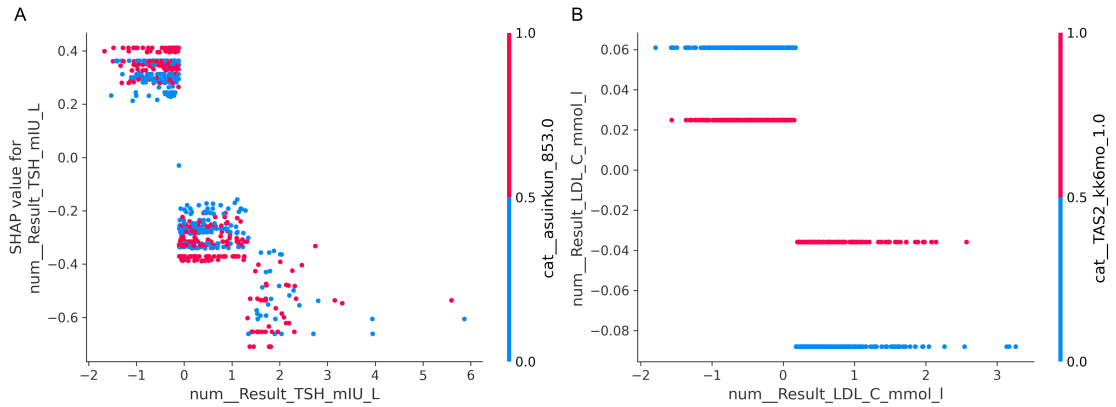

Figure 6.5: Interaction of A) gestational serum TSH levels with mother's municipality of residence at delivery and B) gestational serum LDL levels with the answer '1=not true at all' to the question of feelings 'difficult to find words to describe' (TAS2) at 6 months follow-up. In the y-axis are the SHAP importances for serum biomarkers whereas x-axis shows the actual scaled biomarker levels. The bar on the right hand side with coloring indicates the value of the interacting feature. The dots in the plot indicate individual observations.

In order to determine whether serum biomarker levels have synergy with the other features, SHAP interaction values were computed. According to the results, TSH interacted with features, such as Turku as municipality of residence of the mother at delivery (asuinkun 853.0) (Figure 6.5A), answer totally true to the question of the childhood family and mother's father letting her go out (PBI FA 22 rp2mo), as well as mother's score 5 for impression of the baby being peaceful-easily irritable at scale 1-7 at 3 months follow-up (vaikute1 kk3mo). For example, influence of TSH on total difficulties outcome was stronger for those mother's who lived in Turku and it was weaker for those mother's whose father let them go out in their childhood. LDL interacted with the answer 'not true at all' to the question about feelings 'difficult to find words to describe' (TAS2 kk6mo) at six months follow-up. This answer seemed to decrease the influence of LDL on the outcome (Figure 6.5B).

Mother's rating of 5 at range 1-7 of baby being peaceful - more easily irritable at three months follow-up was associated with decreased risk of total difficulties at five year follow-up. This feature (rating 5) also interacted with the gestational TSH levels. Based on the visual examination (Figure 6.6) or statistical testing there was no correlation with the TSH levels and this feature, although some of the individual mothers in the lower rank groups 1-3 seemed to have higher TSH levels (Figure 6.6 a and d). The total difficulties scores seemed to be higher for those mothers who replied 5-7 to this question (Figure 6.6 b and e, Point-Biserial coefficient 0.14, p-value 0.00002). However, combined examination of these three features (Figure 6.6 c and f) did not reveal any clear patterns. Nevertheless, the interaction between the features in the high dimensional data can be complex and therefore not necessarily clear based on this simplified visualisation.

In addition to serum biomarkers, 94 other features had a Shapley value of at least 0.01. Among the most important ones were for example parents working together once or twice a week (RDAS15), mother's lack of nervousness or mental restlessness, mother age and positive experiences with the baby and own childhood, which all seemed to be protective and decreased the risk of total difficulties at five year follow-up. Child being a boy, mother having fairly little palpitation at the third trimester, and answering partly true to the question about enjoying the discussions with mother of the childhood family, predicted higher risk of total difficulties. Detailed list of the top features and their influence on the outcome is provided in Appendix A (Table A.2).

When children's biomarkers were included in the model, a total of 113 features influenced the outcome with shapley value of at least 0.01. Again gestational TSH

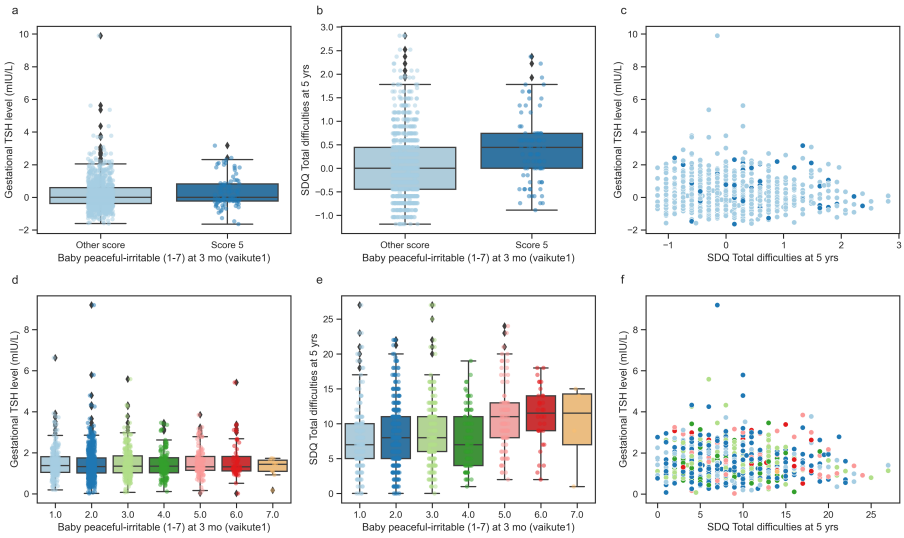

Figure 6.6: Relationships of a) maternal gestational TSH levels and expression/impression of baby being peaceful or easily irritable at range 1-7 at three months follow-up (feature code vaikute1), b) vaikute1 and SDQ Total difficulties score at five year follow-up and c) gestational TSH levels and five year SDQ total difficulties score (color vaikute1) in the entire dataset before (a-c) and after preprocessing (d-e). One outlier with very high TSH value has been removed from the unprocessed data from plot d from category 4.

level was the most important feature (mean 0.20, std 0.05). In addition, LDL (0.06, std 0.03), Ins (0.04, std 0.02), Trigly (0.02, std 0.02) and APOA1 (0.02, std 0.01) were among the important features with weaker influence. The results were therefore consistent with those excluding children's biomarkers from the analysis. However, based on the results, children's own biomarkers were not associated with the SDQ total difficulties outcome.

6.7 Correlation analyses

XGBoost Classifier implements regularisation to control overfitting. Thus, during optimisation and training, it may exclude highly correlated features retaining only one of them in the model. Therefore, as gestational TSH levels were among the most important features predicting total difficulties score, and considering that the features correlating with TSH may have been excluded during modelling, correlation of TSH with the features present in the input dataset, was examined. Correlation analysis was performed with non-parametric Spearman's rank test as well as Point-Bisearial correlation test, which is suitable for comparing continuous numeric variables with binary features. The correlation was tested by using the cleaned and preprocessed input dataset and p-values were corrected for multiple testing with Benjamini-Hochberg method (FDR). According to the results, TSH levels did not have clear statistically significant correlations with other features in the input data.

Correlation of other serum biomarkers measured during pregnancy was also examined to determine, whether they correlate with any of the important features

identified with SHAP test. However, no such correlation were identified. Mother's BMI had weak correlation with the Insulin levels (Spearman's rho 0.23, Benjamini-Hochberg corrected p-value, FDR, 1.82E-12) and modest correlation with CRP (Spearman's rho 0.39, FDR 1.61E-35). Mother's BMI at delivery was among the most important features predicting SDQ 5 year total difficulties outcome, when children's biomarkers were included in the input data (mean Shapley value 0.04, std 0.03). These results together indicate that no informative biomarkers were lost due to penalisation by XGBC.

Finally, correlation between maternal and children's serum biomarkers with BIT-SEA and SDQ scores at both 4 and 5 year follow-ups were examined. No statistically significant correlations were identified. The Spearman's correlation coefficient of TSH and the five year total difficulties SUM score was -0.06 (p-value 0.09).

Details of the computing environment and libraries used in the study are provided in the Appendix B.

7 Discussion

Emerging evidence suggests that disturbances in the prenatal biochemical environment can have adverse effects on the neurodevelopment. However, to validate this link and to identify important factors influencing the outcomes, more large scale studies are needed. This study exploited machine learning to determine whether biomarkers measured from the maternal serum during pregnancy can predict socio-emotional and behavioural outcomes of the children. Furthermore, association of children's own biomarkers with their phenotypes were examined. Machine learning approach with the longitudinal FinnBrain dataset allowed us to control for potential confounders and to identify co-influence of comprehensive feature set collected in the context of the FinnBrain Birth Cohort Study elucidating impact of early life environment on the development and health of children.

According to the results, TSH levels measured from the serum of pregnant mothers at gestational week 24, in combination with a panel of other features, predict SDQ total difficulties of their children at both four and five year follow-ups. Lower gestational TSH levels were associated with increased risk of total difficulties, whereas higher levels were associated with decreased risk. TSH is a clinical biomarker used to detect thyroid dysfunction. Levels of TSH are increased in hypothyroidism

and decreased in hyperthyroidism, whereas T4 hormone (thyroxin) has opposite pattern. What can be considered as a normal TSH level during pregnancy is a matter of current debate, however, in Finnish population, interval 0.07-2.5 mU/l has been recommended as a normal reference range [76]. In the entire FinnBrain dataset 150 women had a value above the range and only 7 below the recommended reference interval. This suggests that even gestational levels of TSH within the reference interval in combination with the other important features can predict later emotional and behavioural problems of the children.

TSH levels predicted behavioural and emotional problems in interaction with other features. The strongest interactions were found with the municipality of residence of the mother at delivery, answer totally true to the question of the childhood family and mother's father letting her go out, and mother's impression of the baby being peaceful-easily irritable at scale. Validity and significance of these interactions requires further studies. Based to the results, living in city of Turku increased the influence of TSH on the total difficulties outcome. Assuming that the finding can be generalised beyond this dataset, one possible explanation can be that the living environment and life style in the bigger city in comparison to the smaller municipalities may influence the maternal hormone levels during pregnancy and increase the risk of total difficulties of the children. The serum biomarker measurements were performed at the same unit and therefore are not explainable by the technical variation between laboratories.

Also several other serum biomarkers were found among the important features. These included LDL, APOA1, Trigly, FT4, Glucose HK2 and insulin. LDL had strong correlation with APOB and cholestrol levels indicating that the influence of

these features on the model can be similar or redundant. According to the results decreased levels of LDL, APOA1, Trigly, FT4 were associated with increased risk of total difficulties and vice versa, whereas higher glucose HK2 levels levels predicted increased and low insulin levels associated with decreased risk of total difficulties. The results suggest that imbalance between glucose and lipid and thyroid hormone metabolism may be associated with the adverse outcome. However, the influence of all the other biomarkers in comparison to TSH on the model was weaker. Further studies are needed to define the physiological relevance of these findings.

Our results on the importance of thyroid hormones are consistent with numerous previous studies reporting association of these hormones with socio-emotional and behavioural problems. However, the directionality in our study differs from many of the previous findings. For example, higher levels of TSH during pregnancy have been linked to higher externalising scores in both genders [43], ADHD symptoms in girls [83] and during first trimester with attention problems in boys [37]. Also subclinical hypothyroidism (increased TSH and normal FT4) at mean gestational weeks 14 predicts increased ADHD symptoms in four year old children, and oppositional-defiant and conduct problems in 6 year old children [56]. Furthermore, high and non-increasing TSH levels in combination with low FT4 levels throughout the pregnancy have been linked to symptoms of anxiety and depression in early childhood [37]. Low levels of FT4 during first and second trimesters of pregnancy have also been associated with increased risk of autism [92], ADHD symptoms [78] and schizophrenia [47]. Consistently, also in our study the lower FT4 levels were associated with higher risk of total difficulties, although pattern would perhaps be expected to be opposite to TSH levels. However, complex in-

teractions may exist between numerous features present in the data and therefore interpretation, and in particular comparison of the findings to other studies, may not be straightforward.

Conversely, also several studies exist which have not found any link between maternal thyroid function and child's neurodevelopment. For example, Fetene et al. (2020) reported null findings for the association between thyroid hormone levels measured during the first trimester of pregnancy (N=4839) and SDQ outcomes of the children at the ages of 3.5, 6.75, 9 or 11 years [42]. Similarly, Chevrier et al. 2011 found no relationship between second trimester TSH levels and socio-emotional or behavioural problems. Yet, perhaps more in line with our findings, they did reported positive association between TSH levels and temporary performance of children in cognition and language at one year follow-up, and improved attention at the age of five years [31].

Nevertheless, the role of thyroid hormones in brain development and function is known to be indispensable. Thyroid hormones are required for generation and differentiation of neural cells, regulate migration, development of synapses, myelination and brain structures in time dependent manner. These hormones are also crucial in maintenance of brain functions throughout the life. During early prenatal development the fetus relies solely on the maternal hormone reservoirs, until it's own endogenous production is launched. However, maternal hormones can be detected in the fetus also at term and deficient thyroid hormone signalling can lead to permanent alterations in the neuronal functions and neurological disorders. [14]. Therefore, our findings are in good agreement with the current understanding in this field.

In our data also high glucose levels predicted total difficulties at five year follow-up. This finding is consistent with previous studies. Elevated glucose levels and insulin resistance are hallmarks of gestational diabetes. Gestational diabetes again has been repeatedly associated with adverse neurodevelopmental outcomes affecting cognition, language development, attention, impulsivity and behaviour of the children [85]. Gestational glucose levels have been also associated with SDQ externalising scores of three and five year olds [38] and risk of conduct problems at 4-16 year olds [63].

Not many studies on maternal gestational serum lipids and child behavioural and emotional outcomes were found in the literature. The study by Kwok et al. found link between lower gestational HDL levels and decreased hyperactivity and higher triglyserides with increased hyperactivity problems [63]. Our directionality on total difficulties was opposite, lower levels of these lipids (considering correlations) were associated with increased risk of difficulties and vice versa. Pinho et al measured LDL levels from adolescents own serum samples and found reduced LDL levels to associated with ADHD based on assessment with SDQ hyperactivity-inattention subscale. [87]. We did not find association between any of the biomarkers measured from children's own serum and emotional and behavioural problems. However, we cannot exclude the possibility that the sample size was not large enough for machine learning approach.

The data mining aim of this study was to develop machine learning models which can be used to predict the outcomes of interest with acceptable performance. Several algorithms and approaches were compared to find the most optimal solution. Consistently with its competative performance in previous studies [30, 117, 13,

94], XGBC was found to be the most robust algorithm also for this dataset. When no features with missing values were allowed in the input data the models did not perform well. This indicates that important features that contribute to the model performance were lost. Similar performances were achieved with input data containing either 20 % or 50 % of missing values.

Training of the model, however, was challenging. The data was high-dimensional and number of observations was rather low. Due to inherent nature of the scoring system, the outcome features had skewed distributions. After trying several different approaches, the best performance was achieved by using XGBoost Classifier and by assigning weights for the training observations belonging to the majority class based on their importance. With this approach moderate validation and generalisation performances were achieved for the models predicting SDQ outcomes. This is logical, since the questionnaire studies have been initially developed to identify individuals with potential problems rather diagnose or classify the children accurately to different categories. The scores are also probably noisy and can be affected by several factors, such as population, ethnic background, assessor, age and in some cases gender [101, 81]. Moreover, the cutoffs in this study were adjusted based on literature and may not have been fully optimal, as standardised cutpoints have not been established for Finnish population. Future studies would most likely benefit of thresholding and harmonisation of the cutpoints. In addition, machine learning from imbalanced datasets seems to be an area of active research. Further studies will probably reveal more effective solutions for challenging datasets with skewed distributions and small number of minority class examples. Nevertheless, although the performance level of models predicting SDQ outcomes remained at

moderate level in this study, similar performances have been previously described in the literature for other health [12] and neuropsychiatric outcomes [70, 103, 65].

The models predicting BITSEA did not achieve acceptable performance level and were rejected. There can be several reasons for the poor performance. One possibility is that the input data did not include features that would explain the outcome, or the time point of assessment was too early to detect the influence of risk factors. Alternatively, similarly to the other outcomes, it is possible that the data was not sufficient in size or was too noisy for the model hampering the learning of the informative features from the data. It is also possible that the threshold used to classify the children for modelling were not optimal. Therefore, future studies would probably benefit of standardisation of the cutpoints for Finnish population.

As a summary, our results suggest that the total difficulties of the children at five year follow-up are influenced by imbalance in thyroid, lipid and glucose metabolism in cross-talk with numerous other features during pregnancy and early life environment. Comparison of our findings to the previous research revealed that the results across studies are variable. However, so are the designs and applied methods, which complicates the conclusions. Nevertheless, consistent finding is that the components of thyroid, glucose and lipid metabolism associate repeatedly with behavioural and emotional problems. Careful review of the similarities and differences between these studies, including variables, timepoints, sample sizes, statistical models and treatment of outcome factors would facilitate understanding of the commonalities and discrepancies in findings. Identification of the important covariates co-influencing the outcomes can help to harmonise datasets and replicate findings across different cohorts in future studies.

Although several interesting factors predicting emotional and behavioural problems of the children at five years of age were identified, the causal relationship cannot be established based on these results. It is possible that other hidden factors, such as shared genetic risk for thyroid dysfunction and behavioural and emotional disorders have influenced the outcomes. For example, Soheili et al. 2023 reported strong genetic correlation between thyroid disease and major depressive disorder as well as with anxiety disorder. However, they did not found genetic correlation between mood disorders and TSH or FT4 levels. [100]. Based on our results we are not able to exclude the possibility that the risk emotional and behavioural problems in early childhood is not mediated by both prenatal biochemical environment and early life social environment. However, the potential causal link remains to be elucidated in further studies.

8 Conclusions

This study identified gestational TSH levels among the most important features predicting the emotional and behavioural problems of the children at both four and five year follow-ups. TSH was not critical for the model performance, however, together with the other features had strong influence on the outcome. Also several other serum biomarkers, including LDL, APOA1, Trigly, FT4, Glucose HK2 and insulin, predicted the five year outcome, however, only at five years follow-up and with much weaker influence. In addition, numerous other protective and risk factors predicting the outcome were identified. No gestational serum biomarkers predicting BITSEA Problem or Competence outcomes at two year follow-up were found. Children's own biomarkers measured at five year follow-up were also not associated with the emotional and behavioural problems. Our results suggest that imbalance in maternal thyroid, lipid and glucose metabolism in cross-talk with numerous other prenatal and early-life factors influence the total difficulties outcome of the children at five year follow-up. These results are important in advancing our understanding of the early life factors associated with emotional and behavioural problems in the childhood and provide predictive markers for early detection of individuals at risk.

References

[1] Md Manjurul Ahsan et al. "Effect of data scaling methods on machine learning algorithms and model performance". In: *Technologies* 9.3 (2021), p. 52.

[2] Jaana Alakortes et al. "Finnish mothers' and fathers' reports of their boys and girls by using the Brief Infant-Toddler Social and Emotional Assessment (BITSEA)". In: *Infant Behavior and Development* 39 (2015), pp. 136–147.

[3] E Alberta. "Community university partnership for the study of children youth and families". In: *Review of the Brief Infant-Toddler Social and Emotional Assessment (BITSEA). Edmonton, Alberta, Canada.* 2011.

[4] Lucas BV de Amorim, George DC Cavalcanti, and Rafael MO Cruz. "The choice of scaling technique matters for classification performance". In: *Applied Soft Computing* 133 (2023), p. 109924.

[5] Stine Linding Andersen, Jørn Olsen, and Peter Laurberg. "Maternal thyroid disease in the Danish National Birth Cohort: prevalence and risk factors". In: *European Journal of Endocrinology* 174.2 (2016), pp. 203–212.

[6] Stine Linding Andersen et al. “Attention deficit hyperactivity disorder and autism spectrum disorder in children born to mothers with thyroid dysfunction: a D anish nationwide cohort study”. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 121.11 (2014), pp. 1365–1374.

[7] Hjördis Ó Atladóttir et al. “Maternal infection requiring hospitalization during pregnancy and autism spectrum disorders”. In: *Journal of autism and developmental disorders* 40 (2010), pp. 1423–1430.

[8] Maria E Baardman et al. “The role of maternal-fetal cholesterol transport in early fetal life: current insights”. In: *Biology of reproduction* 88.1 (2013), pp. 24–1.

[9] Susana Barbosa et al. “Immune activity at birth and later psychopathology in childhood”. In: *Brain, Behavior, & Immunity-Health* 8 (2020), p. 100141.

[10] Susana Barbosa et al. “Serum cytokines associated with behavior: a cross-sectional study in 5-year-old children”. In: *Brain, behavior, and immunity* 87 (2020), pp. 377–387.

[11] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. “A study of the behavior of several methods for balancing machine learning training data”. In: *ACM SIGKDD explorations newsletter* 6.1 (2004), pp. 20–29.

[12] Gopi Battineni et al. “Applications of machine learning predictive models in the chronic disease diagnosis”. In: *Journal of personalized medicine* 10.2 (2020), p. 21.

[13] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms". In: *Artificial Intelligence Review* 54 (2021), pp. 1937–1967.

[14] Juan Bernal. *Thyroid Hormones in Brain Development and Function.* Endotext, 2000-2022.

[15] Michael R Berthold et al. *Guide to intelligent data analysis: how to intelligently make sense of real data*. Springer Science & Business Media, 2010.

[16] Qifang Bi et al. "What is machine learning? A primer for the epidemiologist". In: *American journal of epidemiology* 188.12 (2019), pp. 2222–2239.

[17] Anne-Mari Borg et al. "Finnish norms for young children on the Strengths and Difficulties Questionnaire". In: *Nordic Journal of Psychiatry* 68.7 (2014), pp. 433–442.

[18] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". In: *PloS one* 12.6 (2017), e0177678.

[19] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[20] Margaret J Briggs-Gowan and Alice S Carter. "Social-emotional screening status in early childhood predicts elementary school outcomes". In: *Pediatrics* 121.5 (2008), pp. 957–962.

[21] Margaret J Briggs-Gowan et al. "The Brief Infant-Toddler Social and Emotional Assessment: screening for social-emotional problems and delays in competence". In: *Journal of pediatric psychology* 29.2 (2004), pp. 143–155.

[22] MJ Briggs-Gowan and AS Carter. "BITSEA: Brief Infant-Toddler Social and Emotional Assessment (Examiner's manual) San Antonio". In: *TX: Harcourt Assessment* (2006).

[23] Gürol Canbek, Tugba Taskaya Temizel, and Seref Sagiroglu. "BenchMetrics: A systematic benchmarking method for binary classification performance metrics". In: *Neural Computing and Applications* 33.21 (2021), pp. 14623–14650.

[24] Ilona Carneiro. *Introduction to Epidemiology: Understanding Public Health.* eng. 3rd ed. Maidenhead: McGraw-Hill Education, 2017. ISBN: 0335243185.

[25] Rich Caruana and Alexandru Niculescu-Mizil. "An empirical comparison of supervised learning algorithms". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 161–168.

[26] Tianfeng Chai and Roland R Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature". In: *Geoscientific model development* 7.3 (2014), pp. 1247–1250.

[27] Peter Chapman. "CRISP-DM 1.0: Step-by-step data mining guide". In: 2000. URL: `https://api.semanticscholar.org/CorpusID:59777418`.

[28] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[29] NV Chawla, N Japkowicz, and A Kotcz. *Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explor Newsl. 2004; 6 (1): 1–6.*

[30] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[31] Jonathan Chevrier et al. "Maternal thyroid function during the second half of pregnancy and child neurodevelopment at 6, 12, 24, and 60 months of age". In: *Journal of thyroid research* 2011 (2011).

[32] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1 (2020), pp. 1–13.

[33] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation". In: *PeerJ Computer Science* 7 (2021), e623.

[34] Roshan Chudal et al. "Maternal serum C-reactive protein (CRP) and offspring attention deficit hyperactivity disorder (ADHD)". In: *European child & adolescent psychiatry* 29 (2020), pp. 239–247.

[35] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20 (1995), pp. 273–297.

[36] Artuur Couckuyt et al. "Challenges in translational machine learning". In: *Human Genetics* 141.9 (2022), pp. 1451–1466.

[37] Joyce J Endendijk et al. "Maternal thyroid hormone trajectories during pregnancy and child behavioral problems". In: *Hormones and behavior* 94 (2017), pp. 84–92.

[38] Sabrina Faleschini et al. “Maternal Hyperglycemia in Pregnancy and Offspring Internalizing and Externalizing Behaviors”. In: *Maternal and Child Health Journal* (2023), pp. 1–9.

[39] Manuel Fernández-Delgado et al. “An extensive experimental survey of regression methods”. In: *Neural Networks* 111 (2019), pp. 11–34.

[40] Manuel Fernández-Delgado et al. “Do we need hundreds of classifiers to solve real world classification problems?” In: *The journal of machine learning research* 15.1 (2014), pp. 3133–3181.

[41] Dagnachew Muluye Fetene, Kim S Betts, and Rosa Alati. “Maternal thyroid dysfunction during pregnancy and behavioural and psychiatric disorders of children: a systematic review.” In: *European Journal of Endocrinology* 177.5 (2017).

[42] Dagnachew Muluye Fetene et al. “Maternal prenatal thyroid function and trajectories of offspring emotional and behavioural problems: findings from the ALSPAC cohort”. In: *European Child & Adolescent Psychiatry* 29 (2020), pp. 871–879.

[43] Akhgar Ghassabian et al. “Maternal thyroid function during pregnancy and behavioral problems in the offspring: the generation R study”. In: *Pediatric research* 69.7 (2011), pp. 454–459.

[44] Ivy Giserman Kiss et al. “Developing autism screening criteria for the brief infant toddler social emotional assessment (BITSEA)”. In: *Journal of autism and developmental disorders* 47 (2017), pp. 1269–1277.

[45] Mary Margaret Gleason et al. “Addressing early childhood emotional and behavioral problems”. In: *Pediatrics* 138.6 (2016).

[46] Anna Goodman and Robert Goodman. “Strengths and difficulties questionnaire as a dimensional measure of child mental health”. In: *Journal of the American Academy of Child & Adolescent Psychiatry* 48.4 (2009), pp. 400–403.

[47] David Gyllenberg et al. “Hypothyroxinemia during gestation and offspring schizophrenia in a national birth cohort”. In: *Biological Psychiatry* 79.12 (2016), pp. 962–970.

[48] Helena Haapsamo et al. “Screening infants with social and emotional problems: A pilot study on the brief infant oddler social and emotional assessment (Bitsea) in northern Finland”. In: *International journal of circumpolar health* 68.4 (2009), pp. 386–393.

[49] Velda X Han et al. “Maternal immune activation and neuroinflammation in human neurodevelopmental disorders”. In: *Nature Reviews Neurology* 17.9 (2021), pp. 564–579.

[50] David J Hand. “Intelligent data analysis: Issues and opportunities”. In: *International Symposium on Intelligent Data Analysis*. Springer. 1997, pp. 1–14.

[51] Luke S Heuer et al. “An exploratory examination of neonatal cytokines and chemokines as predictors of autism risk: the early markers for autism study”. In: *Biological psychiatry* 86.4 (2019), pp. 255–264.

[52] Anne AH de Hond, Ewout W Steyerberg, and Ben van Calster. "Interpreting area under the receiver operating characteristic curve". In: *The Lancet Digital Health* 4.12 (2022), e853–e855.

[53] Ghulam Hussain et al. "Role of cholesterol and sphingolipids in brain development and neurological diseases". In: *Lipids in health and disease* 18.1 (2019), pp. 1–12.

[54] Karen L Jones et al. "Autism with intellectual disability is associated with increased levels of maternal cytokines and chemokines during gestation". In: *Molecular psychiatry* 22.2 (2017), pp. 273–279.

[55] Miia Kaartinen et al. "Maternal tiredness and cytokine concentrations in mid-pregnancy". In: *Journal of Psychosomatic Research* 127 (2019), p. 109843.

[56] Mariza Kampouri et al. "Maternal mild thyroid dysfunction and child behavioral and emotional difficulties at 4 and 6 years of age: the Rhea Mother-Child Cohort study, Crete, Greece". In: *Hormones and behavior* 116 (2019), p. 104585.

[57] Koray Karabekiroglu et al. "The clinical validity and reliability of the Brief Infant–Toddler Social and Emotional Assessment (BITSEA)". In: *Infant Behavior and Development* 33.4 (2010), pp. 503–509.

[58] Linnea Karlsson et al. "Cohort profile: the FinnBrain birth cohort study (FinnBrain)". In: *International journal of epidemiology* 47.1 (2018), 15–16j.

[59] Linnea Karlsson et al. "Cytokine profile and maternal depression and anxiety symptoms in mid-pregnancy—the FinnBrain Birth Cohort Study". In: *Archives of women's mental health* 20 (2017), pp. 39–48.

[60] Paula Krakowiak et al. "Neonatal cytokine profiles associated with autism spectrum disorder". In: *Biological psychiatry* 81.5 (2017), pp. 442–451.

[61] Ingrid Kruizinga et al. "Screening accuracy and clinical application of the Brief Infant-Toddler Social and Emotional Assessment (BITSEA)". In: *PloS one* 8.8 (2013), e72602.

[62] Ajay Kulkarni, Deri Chong, and Feras A Batarseh. "Foundations of data imbalance and solutions for a data democracy". In: *Data democracy*. Elsevier, 2020, pp. 83–106.

[63] Janell Kwok et al. "Examining maternal cardiometabolic markers in pregnancy on child emotional and behavior trajectories: Using growth curve models on a cohort study". In: *Biological Psychiatry Global Open Science* 3.4 (2023), pp. 614–622.

[64] Timothy L Lash et al. *Modern Epidemiology*. eng. 4th ed. Philadelphia: Wolters Kluwer Health, 2021. ISBN: 9781451193282.

[65] John V Lavigne, Kathryn Mendelsohn Meyers, and Marissa Feldman. "Systematic review: Classification accuracy of behavioral screening measures for use in integrated primary care settings". In: *Journal of pediatric psychology* 41.10 (2016), pp. 1091–1109.

[66] Brian K Lee et al. "Maternal hospitalization with infection during pregnancy and risk of autism spectrum disorders". In: *Brain, behavior, and immunity* 44 (2015), pp. 100–105.

[67] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Ma-

chine Learning". In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[68] Marika Leppänen et al. "Burden of mental, behavioral, and neurodevelopmental disorders in the Finnish most preterm children: a national register study". In: *European Child & Adolescent Psychiatry* (2023), pp. 1–8.

[69] Qingyun Li and Ben A Barres. "Microglia and macrophages in brain homeostasis and disease". In: *Nature Reviews Immunology* 18.4 (2018), pp. 225–242.

[70] Maria Chiara Liverani et al. "Behavioral outcome of very preterm children at 5 years of age: Prognostic utility of brain tissue volumes at term-equivalent-age, perinatal, and environmental factors". In: *Brain and Behavior* 13.2 (2023), e2818.

[71] Riikka J Lund et al. "Placental DNA methylation marks are associated with maternal depressive symptoms during early pregnancy". In: *Neurobiology of Stress* 15 (2021), p. 100374.

[72] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[73] Scott M Lundberg et al. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[74] Cecilie N Lydholm et al. "Parental infections before, during, and after pregnancy as risk factors for mental disorders in childhood and adolescence: a

nationwide Danish study". In: *Biological Psychiatry* 85.4 (2019), pp. 317–325.

[75] Erika M Manczak and Ian H Gotlib. "Lipid profiles at birth predict teacher-rated child emotional and social development 5 years later". In: *Psychological science* 30.12 (2019), pp. 1780–1789.

[76] Tuija Männistö et al. "Early pregnancy reference intervals of thyroid hormone concentrations in a thyroid antibody-negative pregnant population". In: *Thyroid* 21.3 (2011), pp. 291–298.

[77] Yoshihiro Miyake et al. "Maternal fat intake during pregnancy and behavioral problems in 5-y-old Japanese children". In: *Nutrition* 50 (2018), pp. 91–96.

[78] Thiago Modesto et al. "Maternal mild thyroid hormone insufficiency in early pregnancy and attention-deficit/hyperactivity disorder symptoms in children". In: *JAMA pediatrics* 169.9 (2015), pp. 838–845.

[79] Jason Denzil Morgenstern et al. "Predicting population health with machine learning: a scoping review". In: *BMJ open* 10.10 (2020), e037860.

[80] Rashmi Mullur, Yan-Yun Liu, and Gregory A Brent. "Thyroid hormone regulation of metabolism". In: *Physiological reviews* (2014).

[81] Carsten Obel et al. "The strengths and difficulties questionnaire in the Nordic countries". In: *European child & adolescent psychiatry* 13 (2004), pp. ii32–ii39.

[82] Michael O Ogundele and Michael Morton. "Classification, prevalence and integrated care for neurodevelopmental and child mental health disorders: A brief overview for paediatricians". In: *World journal of clinical pediatrics* 11.2 (2022), p. 120.

[83] Fanni Päkkilä et al. "The impact of gestational thyroid hormone concentrations on ADHD symptoms of the child". In: *The Journal of Clinical Endocrinology & Metabolism* 99.1 (2014), E1–E8.

[84] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[85] Robert Perna et al. "Gestational diabetes: long-term central nervous system developmental and cognitive sequelae". In: *Applied Neuropsychology: Child* 4.3 (2015), pp. 217–220.

[86] Jianmin Piao et al. "Alarming changes in the global burden of mental disorders in children and adolescents from 1990 to 2019: a systematic analysis for the Global Burden of Disease study". In: *European Child & Adolescent Psychiatry* 31.11 (2022), pp. 1827–1845.

[87] Raquel Pinho et al. "Attention-deficit/hyperactivity disorder is associated with reduced levels of serum low-density lipoprotein cholesterol in adolescents. Data from the population-based German KiGGS study". In: *The World Journal of Biological Psychiatry* 20.6 (2019), pp. 496–504.

[88] Foster Provost. "Machine learning from imbalanced data sets 101". In: *Proceedings of the AAAI'2000 workshop on imbalanced data sets*. Vol. 68. 2000. AAAI Press. 2000, pp. 1–3.

[89] Olli Rajasilta et al. “Maternal pre-pregnancy BMI associates with neonate local and distal functional connectivity of the left superior frontal gyrus”. In: *Scientific Reports* 11.1 (2021), p. 19182.

[90] Sebastian Raschka. “An overview of general performance metrics of binary classifier systems”. In: *arXiv preprint arXiv:1410.5330* (2014).

[91] Jerod M Rasmussen et al. “Maternal pre-pregnancy body mass index is associated with newborn offspring hypothalamic mean diffusivity: a prospective dual-cohort study”. In: *BMC medicine* 21.1 (2023), pp. 1–13.

[92] Gustavo C Román et al. “Association of gestational maternal hypothyroxinemia and increased autism risk”. In: *Annals of neurology* 74.5 (2013), pp. 733–742.

[93] Julie B Rosenberg et al. “Maternal inflammation during pregnancy is associated with risk of ADHD in children at age 10”. In: *Brain, Behavior, and Immunity* (2023).

[94] Iqbal H Sarker. “Machine learning: Algorithms, real-world applications and research directions”. In: *SN computer science* 2.3 (2021), p. 160.

[95] Steven Schepanski et al. “Prenatal immune and endocrine modulators of offspring’s brain development and cognitive functions later in life”. In: *Frontiers in immunology* 9 (2018), p. 2186.

[96] *SDQ info*. https://www.sdqinfo.org/. Accessed: 2023-12-02.

[97] Bridgette D Semple et al. “Brain development in rodents and humans: Identifying benchmarks of maturation and vulnerability to injury across species”. In: *Progress in neurobiology* 106 (2013), pp. 1–16.

[98] John Shawe-Taylor and Nello Cristianini. "Margin distribution and soft margin". In: *Advances in Large Margin Classifiers* (2000), pp. 349–358.

[99] Tanin Sirimongkolkasem and Reza Drikvandi. "On regularisation methods for analysis of high dimensional data". In: *Annals of Data Science* 6 (2019), pp. 737–763.

[100] Sourena Soheili-Nezhad et al. "Exploring the genetic link between thyroid dysfunction and common psychiatric disorders: A specific hormonal or a general autoimmune comorbidity". In: *Thyroid* 33.2 (2023), pp. 159–168.

[101] Lisanne L Stone et al. "Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4-to 12-year-olds: a review". In: *Clinical child and family psychology review* 13 (2010), pp. 254–274.

[102] Meagan R Talbott and Meghan R Miller. "Future directions for infant identification and intervention for autism spectrum disorder from a transdiagnostic perspective". In: *Journal of Clinical Child & Adolescent Psychology* 49.5 (2020), pp. 688–700.

[103] Ashley E Tate et al. "Predicting mental health problems in adolescence using machine learning techniques". In: *PloS one* 15.4 (2020), e0230389.

[104] Alberto Traverso et al. "Diving deeper into models". In: *Fundamentals of clinical data science* (2019), pp. 121–133.

[105] *UCI datasets*. https://archive.ics.uci.edu/datasets/. Accessed: 2023-12-03.

[106] Miraç Barış Usta and Koray Karabekiroğlu. "Does the Psychopathology of the Parents Predict the Developmental-Emotional Problems of the Toddlers?" In: *Archives of Neuropsychiatry* 57.4 (2020), p. 265.

[107] Erin R Wallace et al. "Prenatal urinary metabolites of polycyclic aromatic hydrocarbons and toddler cognition, language, and behavior". In: *Environment International* 159 (2022), p. 107039.

[108] Ni Wayan Surya Wardhani et al. "Cross-validation metrics for evaluating classification performance on imbalanced data". In: *2019 international conference on computer, control, informatics and its applications (IC3INA)*. IEEE. 2019, pp. 14–18.

[109] Vitor Werner de Vargas et al. "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study". In: *Knowledge and Information Systems* 65.1 (2023), pp. 31–57.

[110] Cort J Willmott. "Some comments on the evaluation of model performance". In: *Bulletin of the American Meteorological Society* 63.11 (1982), pp. 1309–1313.

[111] Dennis L Wilson. "Asymptotic properties of nearest neighbor rules using edited data". In: *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972), pp. 408–421.

[112] David H Wolpert and William G Macready. "No free lunch theorems for optimization". In: *IEEE transactions on evolutionary computation* 1.1 (1997), pp. 67–82.

[113] X Wu et al. "Top 10 algorithms in data mining Knowledge and Information Systems, vol. 14, no. 1". In: *Dec* (2007).

[114] *XGboost documentation*. https://xgboost.readthedocs.io/en/stable/tutorials/model.html. Accessed: 2023-12-18.

[115] Fuliang Yi et al. "XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease". In: *BMC Medical Informatics and Decision Making* 23.1 (2023), p. 137.

[116] Kristine E Zengeler and John R Lukens. "Innate immunity at the crossroads of healthy brain maturation and neurodevelopmental disorders". In: *Nature Reviews Immunology* 21.7 (2021), pp. 454–468.

[117] Chongsheng Zhang et al. "An up-to-date comparison of state-of-the-art classification algorithms". In: *Expert Systems with Applications* 82 (2017), pp. 128–150.

[118] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.

Appendix A Features predicting SDQ total difficulties.

Table A.1: Features predicting SDQ total difficulties at four year follow-up identified with SHAP technique.

Feature	Description	Timepoint	Response (range)	Risk of difficulties	SHAP mean	SHAP std
TAS7_kk6	feelings7, puzzled with feelings of the body	6 mo	1=not true at all (1-5)	decrease	0.30	2.23E-02
BNSQ2b_rp3	sleeping2, awake minute leisure	gwk 34	continuous	increase	0.20	3.61E-02
PBQ25_kk3	relation to the baby25, baby easily consoled	3 mo	5=very often (1-6)	decrease	0.16	1.58E-02
SPSQ28_kk3	experiences of the parenthood33, positive attitude to the life	3 mo	5=very true (1-5)	decrease	0.15	2.00E-02
PRAQsum10_rp2	PRAQ sum score (anxiety symptoms during pregnancy)	gwk 24	continuous	increase	0.11	1.14E-02
MDAS4_rp1	dentalcare4, removal of the tartar	gwk 24	4=I would be a little nervous (1-5)	increase	0.11	1.61E-02
SCL10_kk3	anxiety20, nervousness or mental restlessness	3 mo	1=not at all (1-5)	decrease	0.10	1.68E-03
SPSQ9_kk3	exper. of the parenth.14, the age mates not delighted of company	3 mo	1=not true (1-5)	decrease	0.10	1.62E-03
per4_rp1	number of marriages	gwk 14	continuous	increase	0.09	2.57E-02
per2_rp1	beginning year of the present marriage	gwk 14	year	increase	0.07	3.04E-02
SCL5_rp2	anxiety15, tension	gwk 24	1=not at all (1-5)	decrease	0.07	1.35E-02
SCL10_kk6	anxiety10, nervousness or mental restlessness	6 mo	1=not at all (1-5)	decrease	0.07	8.64E-03
ECR15_rp2	human relations, would not like if really knew	gwk 24	1=it does not describe me at all (1-7)	decrease	0.07	1.26E-02
SPSQ32_kk3	experiences of the parenthood37, feels like doesn't manage	3 mo	1=not true (1-5)	decrease	0.06	4.44E-03
TADS35c_rp1	life events35, 13-18 years, friends	gwk 24	4=extremely often (0-4)	decrease	0.06	7.49E-03
vaikute8_kk3	impression and experience of the baby, easy/difficult to interpret	3 mo	2 (1-7)	decrease	0.06	1.04E-02
TADS3b_rp1	life events3, 7-12 years, was bullied	gwk 24	0=never (0-4)	decrease	0.06	1.02E-02
PPRFQ20_rp3	baby and parenthood20, behaves well	gwk 34	1=no, I totally disagree (1-7)	decrease	0.05	1.38E-02
BNSQ5_kk6	sleeping5, too early	6 mo	1=not even once... (1-5)	decrease	0.05	3.79E-03
vaikute6_kk3	impression and experience of the baby, strong frail	3 mo	2 (1-7)	increase	0.05	3.83E-03
vointi7a_kk3	difficulties, eating of the child/feeding	3 mo	2=a little (1-3)	increase	0.04	8.81E-03
hassle2a_rp3	worry2, business	gwk 34	-1=fairly little (-3-0)	decrease	0.04	8.79E-03
BNSQ2a_rp2	sleeping2, minutes awake on working day	gwk 24	continuous	increase	0.04	6.66E-03
TSH_mIU_L	gestational serum TSH levels	gwk 24	continuous	decrease	0.04	5.71E-03
PoPRFQ19_kk6	baby and parenthood19 behaving in a confusing way	6 mo	NA	decrease	0.04	2.71E-03
SPSQ28_kk3	exper. of the parenth.33, positive attitude to the life	3 mo	4 (1=not true . . . 5=very true)	increase	0.04	4.64E-03
AIS7_rp3	sleeping31, ability to function	gwk 34	1=reduced a little (0-3)	decrease	0.04	1.59E-03
vointi7a_kk3	difficulties, eating of the child/feeding	3 mo	1=not any 2=a little 3=considerably	decrease	0.04	8.57E-04
TADS11b_rp1	life events11, 7-12 years, to confide	gwk 24	4=extremely often (0-4)	decrease	0.04	8.43E-04
asuinkun_853.0	Municipality at the birth moment.	birth	853	decrease	0.04	2.34E-03
ika_taytpvm_kk6	The child's age when answering the form (days)	6 mo	continuous	increase	0.04	3.19E-03
TADS3c_rp1	life events3, 13-18 years, was bullied	gwk 24	0=never (0-4)	decrease	0.04	6.77E-03

Table A.2: Features predicting total difficulties at five year follow-up identified using SHAP technique. Features with mean importance of $\geq 0.05 are shown$.

Feature	Description	Timepoint	Response (range)	Risk of difficulties	SHAP mean	SHAP std
TSH_mIU_L	gestational serum TSH level	gwk 24	continuous	decrease	0.34	0.10
RDAS15_rp3	relationship15, working together	gwk 34	4=once/twice per wk (1-6)	decrease	0.21	0.07
SCL10_rp1	anxiety20, nervousness or mental restlessness	gwk 14	1=not at all (1-5)	decrease	0.19	0.07
gender	gender	birth	1=boy (1, 2=girl)	increase	0.17	0.03
TAS11_kk6	feelings11, difficult to describe feelings that the others resonate	6 mo	1=not true at all (1-5)	decrease	0.17	0.11
SCL4_rp3	anxiety14, palpitations	gwk 34	2=fairly little (1-5)	increase	0.17	0.07
vaikute1_kk3	impression and experience of the baby, peaceful-irritable	3 mo	2 (1-7)	decrease	0.17	0.07
PBI_FA_22_rp2	the childhood family, father22, let me go out	gwk 24	2=partly true (1-4)	decrease	0.16	0.05
PBI_FA_12_rp2	the childhood family, father12, smiled often	gwk 24	1=totally true (1-4)	decrease	0.15	0.08
PBI_MO_11_rp2	the childhood family, mother11, enjoyed the discussions	gwk 24	2=partly true (1-4)	increase	0.14	0.03
PoPRFQ3_kk6	baby and parenthood3, interested how does the baby feel	6 mo	7	decrease	0.13	0.06
SOC13_rp2	experience13, controlling feelings	gwk 24	7=seldom or never (1-7)	decrease	0.13	0.12
IKA_MO_SYNT	The mother's age	birth	continuous	decrease	0.13	0.06
PBI_MO_10_rp2	the childhood family, mother10, invaded my privacy	gwk 24	4=not true at all (1-4)	decrease	0.13	0.05
vointi7a_kk3	difficulties, eating of the child/feeding	3 mo	1=not any (1-3)	decrease	0.12	0.02
TAS2_kk6	feelings2, difficult to find words to describe	6 mo	1=not true at all 1-5)	decrease	0.12	0.06
SMUULIEVI	Other non-medical pain relief in delivery	birth	1=yes (0=no, 1=yes)	increase	0.12	0.04
BNSQ11_rp1	sleeping11, tendency to fall asleep at leisure	gwk 14	1=not even once... (1-5)	decrease	0.11	0.04
CDrisk5_rp1	difficulties5, recovery	gwk 14	5=nearly always true (1-5)	decrease	0.11	0.08
VANHEMMAT	The number of parents participating the study	pregnancy	continuous	increase	0.11	0.14
MFAS18_rp1	devotion relation6, in lap	gwk 14	4=yes (1-5)	decrease	0.10	0.07
SPSQ23_kk3	experiences of the parenthood28, feelings of guilt	3 mo	1=not true (1-5)	decrease	0.10	0.04
RDAS4_rp3	relationship4, behavior	gwk 34	2=almost always agree (1-6)	decrease	0.10	0.03
tau4_7_kk6	speech and communication7, school remedial instruction of parents	6 mo	3=no (1-3)	decrease	0.10	0.03
EPDS8_rp3	mood8, sad or miserable feeling	gwk 34	3=not very often (1-4)	increase	0.10	0.03
PBI_FA_6_rp2	the childhood family, father6, was warm and tender	gwk 24	1=totally true (1-4)	decrease	0.10	0.05
synnytysten_lkm	Number of previous births	previously	continuous	decrease	0.09	0.04
MDAS3_kk3	dental care3, drilling	3 mo	4=I would be a little nervous (1-5)	increase	0.09	0.01
SCL5_rp3	anxiety15, tension	gwk 34	1=not at all (1-5)	increase	0.09	0.03
pai1_rp1	intoxicants1, weeks of pregnancy	gwk 14	continuous	decrease	0.08	0.02
BISQ9_kk6	baby's sleep9, putting baby to sleep in the evenings (hours)	6 mo	continuous	increase	0.08	0.00
PBQ25_kk3	relation to the baby25, baby easily consoled	3 mo	5=very often (1-6)	decrease	0.08	0.04
MFAS21_rp2	devotion relation21, hiccup	gwk 24	2=not (1-5)	increase	0.08	0.00
per21_rp1	music as a hobby or study	gwk 14	1=no (1, 2=yes)	increase	0.08	0.00
PBI_FA_25_rp2	the childhood family, father25, let me dress as I wanted	gwk 24	1=totally true (1-4)	increase	0.08	0.00
EPDS8_rp3	mood8, sad or miserable feeling	gwk 34	4=no, not at all (1-4)	decrease	0.08	0.07
MDAS4_rp1	dentalcare4, removal of the tartar	gwk 14	4=I would be a little nervous (1-5)	increase	0.08	0.01
AIS4_rp2	sleeping25, total amount of the sleep	gwk 24	1=to some extent inadequate(0-3)	decrease	0.08	0.02
PNRO1_LUOK	Classification of postal code	birth	1=Turku (2=Åland, 3=Other)	decrease	0.07	0.02
hassle2a_rp2	worry2, business	gwk 24	-2=fairly much (-3-0)	decrease	0.07	0.04
per2_rp1	beginning year of the present marriage	gwk 14	continuous	increase	0.07	0.05
PPRFQ6_rp3	baby and parenthood6, reactions of own parents	gwk 34	6 (1-7)	decrease	0.07	0.04
OHIP3_rp3	health of the mouth and quality of life3, pain or ache	gwk 34	5=one not at all (1-6)	decrease	0.07	0.04
OHIP3_rp3	health of the mouth and quality of life3, pain or ache	gwk 24	4=very seldom (1-6)	increase	0.06	0.04
SOC5_rp2	experience5, unfairly treated	gwk 24	7=very seldom/never (1-7)	decrease	0.06	0.09
per13_4lk_rp1	Estimated monthly income	gwk 14	1=1500 eur (1-4)	increase	0.06	0.04
LDL_C_ml_l	gestational serum LDL level	gwk 24	continuous	decrease	0.06	0.02
SPSQ24_kk3	experiences of the parenthood29, more tired than ordinary	3 mo	2 (1-5)	decrease	0.06	0.08
PBI_MO_8_rp2	the childhood family, mother8, did not want me to grow to be an adult	gwk 24	3=weakly true (1-4)	increase	0.05	0.02
RDAS10_rp3	relationship10, quarreling	gwk 34	3=sometimes (1-6)	decrease	0.05	0.06
SOC11_rp2	experience11, stating afterwards	gwk 24	5 (1-7)	decrease	0.05	0.06
TADS37c_rp1	life events37, 13-18 years, I succeeded in school	gwk 14	3=often (0-4)	decrease	0.05	0.04
asuinkun	Municipality in which the mother lives at the birth moment	birth	853=Turku	decrease	0.05	0.02
PBI_FA_22_rp2	the childhood family, father22, let me go out	gwk 24	1=totally true (1-4)	increase	0.05	0.03
APO_A1_g_l	gestational serum APOA1 level	gwk 24	continuous	decrease	0.05	0.03
PBQ10_kk3	relation to the baby10, irritates	3 mo	continuous	increase	0.05	0.03
BNSQ15b_rp3	sleeping15, afternoon naps hour	gwk 34	continuous	increase	0.05	0.05
MDAS3_rp3	dental care3, drilling	gwk 34	4=I would be a little nervous (1-5)	decrease	0.05	0.02
Gluc_HK2_ml_l	gestational serum Gluc HK2 level	gwk 24	continuous	increase	0.05	0.03

Appendix B Operating system, computing environment and libraries used in the study.

Operating system

macOS-12.5.1-arm64-arm-64bit, total memory 32.0 GB

Computing environment

python 3.11.0

conda 24.3.0

jupyter notebook 7.0.8

List of used libraries and their versions

imblearn 0.11.0

joblib 1.2.0

matplotlib 3.8.0

numpy 1.26.4

pandas 2.2.1

plotly 5.19.0

scipy 1.12.0

seaborn 0.12.2

shap 0.42.1

sklearn 1.4.2

xgboost 1.7.3