A machine learning approach to understanding depression and anxiety in new mothers

Predicting symptom levels using population-based registry data from a large Norwegian prospective study

Maria Linea Horgen

Thesis submitted for the degree of
Master in Computational Science: Physics
60 credits

Department of Physics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2022

A machine learning approach to understanding depression and anxiety in new mothers

Predicting symptom levels using population-based registry data from a large Norwegian prospective study

Maria Linea Horgen

A machine learning approach to understanding depression and anxiety in new mothers

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

This thesis applies several supervised machine learning algorithms to data from the Norwegian Mother, Father and Child Cohort (MoBa) study. It has a twofold objective: investigate methodological differences in predictive performance across models and identify exposures related to symptom levels of depression and anxiety in new mothers. Depression and anxiety following childbirth can have detrimental consequences for the mother, child and family, and it is recognized as a public health concern. The MoBa study has been the subject of nearly 1000 published articles, and very limited research has applied any form of machine learning to the data. This thesis aims to address this gap by comparing the elastic net, deep neural networks, and gradient boosted regression trees using the XGBoost library with a multiple linear regression model to predict depression and anxiety after birth. The linear model is one of the most commonly used statistical models in the field of psychology.

A total of 41 807 mothers were included in the sample, and we measured levels of depression and anxiety through the mean value of the Hopkins Symptoms Checklist (SCL). The mean SCL score was predicted at 6, 18 and 36 months after birth, using both prenatal and postnatal exposures. We created five different datasets from the features measured in the prenatal period to identify possible scenarios where the various models would be most effective. Our methodological comparison found that the multiple linear regression model consistently produced prediction errors between 6% to 11% higher than the best performing machine learning model at each time point, quantified through the root mean squared error (RMSE). The XGBoost model achieved the best performance on a dataset with features with an absolute correlation above 0.35 with the target variable. The mean error over all three time points was an RMSE equal to 0.3716. Our models identified several established risk factors associated with depression and anxiety, such as aspects related to the mother's work situation, attitudes related to weight gain and self-esteem. They have all been found to impact the levels of depression and anxiety after childbirth. Even though the sophisticated machine learning models exhibited the lowest prediction errors, they showed an inadequate ability to identify individuals with symptoms at a clinical level. This led us not to discourage using linear models in psychological research in favor of machine learning. However, the potential of machine learning in the health care domain is not yet fully realized, and we suggest several improvements addressing some of the experienced shortcomings of this thesis.

Acknowledgements

I would like to thank my three supervisors, Morten Hjorth-Jensen, John Aiken and Nikolai Czajkowski, for providing guidance and valuable feedback during this year. This process would have been hard to navigate without somebody to lean on, and I am grateful for the time you all invested in my work.

Before I started my bachelor's in the physics program at the University of Oslo (UiO), I completed one year of the psychology program at the same university. When I made the switch, I had not envisioned that my master thesis would revolve around anything remotely connected to my first year as a student, and I had never even heard of machine learning. So I have to thank the department of physics and the center for computing in science education for allowing me to explore fields in my academic interest and combine them with the computational skills that I have acquired during my time in the physics program.

I would not have been able to realize this thesis if not for the PROMENTA research center at the department of psychology at UiO. I highly appreciate your willingness to participate in this interdisciplinary cooperation and for believing that I could make contributions to your field of research.

Finally, this year would not have been the same without the presence and support of my friends and family!

Contents

List of Figures

List of Tables

Part I

Introduction&Background

Chapter 1

Introduction

Giving birth and experiencing parenthood is generally viewed as a positive life event in the eyes of society. However, for many women, research has shown that childbirth is associated with several negative outcomes, such as a decline in marital relationship [1] and life satisfaction [2], higher levels of anxiety and depression [3] and some even experience symptoms of post-traumatic stress disorder [4]. Not all of these women experience these changes on a clinical level. However, recognizing the implications these strains can have on the mother is essential in order to provide appropriate and sufficient care in the time following childbirth.

Between 10% to 15% of all new mothers experiences clinical levels of depression after giving birth [5, 6]. This form of depression is called postpartum depression, and it describes depression that onsets in the first 12 months after giving birth [7]. It manifests itself as behavioural changes related to loss of interest, constantly feeling unhappy and signs of fatigue [8]. An even higher prevalence is found for postpartum anxiety, with several studies estimating it to be between 15% to 20% [9, 10]. Postpartum anxiety often has a narrower context than general anxiety disorders and is often associated with excess worrying regarding their abilities as parents [11] and difficulties in controlling the worrying. Despite having a higher prevalence than postpartum depression, postpartum anxiety has received less attention. The two are often comorbid, with one study estimating the comorbidity to be 75% [10]. Comorbidity refers to simultaneously experiencing two or more diseases.

Experiencing elevated levels of depression and anxiety following childbirth is associated with a high number of negative consequences. Postpartum anxiety is related to a decrease in maternal self-confidence [12] and preterm delivery [13], while women suffering from postpartum depression have a significantly elevated risk for maternal suicide [14]. Postpartum anxiety and depression are also associated with long-term negative effects on the child's development and can lead to a delay in mental development [15]. It is not only maternal mental health that can have a long-lasting effect on a child. Studies have shown that the parents' general health status, psychological traits and socioeconomic characteristics all impact their ability to provide care for their children [16]. Developing a better understanding of how health conditions that commonly affect new parents, especially mothers, arise and how to prevent them might significantly influence the long-term well-being of both parents and children.

Given the potentially far-reaching consequences of postpartum depression and anxiety, a substantial amount of research has been dedicated to the topic, mainly using conventional data analysis procedures and binary classification [17–20]. The binary targets are often constructed from continuous measures, and specific threshold values are used to categorize the continuous variable. The practice is not well-accepted within the field, as it is associated with several problems, one being the loss of information [21]. Most of the research has been aimed at postpartum depression, and it has been recognized as a public health problem [22]. Here, we use the word "conventional" to refer to methods commonly found in psychological research.

When considering previous research on the topics, it is important to consider that the

conditions for mothers in the postpartum period vary between countries with respect to maternity leave and general support from the health care system. These differences create contrasting conditions for women in the postpartum period, raising questions about the transferability of research performed in other cultures. There is evidence of a lower prevalence of depression following childbirth in cultures that exhibits more substantial social support for the mothers [23]. A study from 1989 indicated that postpartum depression might be a culturally bound syndrome in western societies, such as the U.S, due to a lack of social and instrumental support for new mothers [24].

Although statistical analyses are conducted within a rich methodological framework in psychology, the field is experiencing an ongoing replication crisis. A study from 2015 found that a large number of replication studies conducted on papers from three recognized psychology journals produced reduced support for the original findings [25]. More specifically, 100 replication studies were conducted, and 97% of the original studies reported statistically significant results. However, merely 36% of the replications produced significant results, creating questions about whether we know as much as we claim to. Incorporating selected key principles from the field of machine learning and the use of larger sample sizes have been proposed as one remedy to this crisis [26, 27], and this is a topic further discussed in chapter 2. Machine learning is a field centered around learning patterns from data to make meaningful predictions on novel data, and the methods associated with the field are often capable of handling complex relationships within data. To combat this replication crisis, a shift from traditional modeling to more sophisticated machine learning models could prove beneficial, as machine learning methods are developed around an increased focus on predictions on unseen data instead of maximizing the goodness-of-fit of a model.

Several meta-studies have revealed that two of the most prevalent analysis techniques in psychological literature are analysis of variance and multiple regression [28]. They are often used in conjunction with a small number of independent variables [27] and low sample sizes, usually ranging from 40 to 120 participants [29]. Machine learning models often have a high number of parameters and thus require a large amount of data to achieve adequate results. There has been an overall increase in sample size in the field of psychology in the last ten years; a study from 2019 investigating 1300 papers from the four top empirical social psychology journals reported that the median sample size increased from 117 in 2011 to 195 in 2018 [30]. Despite this increase, together with a tradition of conducting theory-driven research, the use of machine learning is rarely utilized in psychology [31]. Thus the proposed beneficial shift from conventional modeling to more sophisticated machine learning models has not yet taken root in research aimed at postpartum depression and anxiety.

The use of a small number of independent variables in psychological research contrasts with the big data revolution that is currently happening. This revolution refers to information being collected at record-low times and generally becoming more accessible to the public. It has been leading to the use of ever-larger datasets with high numbers of independent variables [32]. The fact that the majority of knowledge in the field of psychology has been acquired through research with a modest number of independent variables raises questions about the need for high-dimensional data in psychological research. Hence, investigating how different numbers of independent variables affect the same dependent variable is an exciting topic to investigate in psychological research.

This thesis will explore data from the Norwegian Mother, Father and Child Cohort Study (MoBa) together with selected machine learning methods to investigate depression and anxiety in what we coin as the extended postpartum period[1]. The MoBa data contains observations from more than 95 000 women on over 7000 variables. The data were collected between 1999 and 2008, starting from the 17th week of pregnancy until the child was 14 years old [33]. As of March 2022, there have been 973 published articles associated with MoBa, covering an extensive range

[1]The postpartum period is usually the time interval between labour and 12 months after birth [7].

of exposures and outcomes. There have been several important findings relating to different exposures, e.g., paracetamol use during pregnancy is associated with delayed neurodevelopment in children [34], high levels of relationship satisfaction predict lower risk of maternal infections during pregnancy [35] and grand-maternal smoking is associated with increased risk of asthma [36].

Given the volume of data gathered in MoBa, analyses through machine learning appear to be a viable choice when investigating statistical relationships in the data. However, to the best of our knowledge, only one out of the 973 published publications [37] describe any use of more sophisticated machine learning methods. The focus of this article was on autism spectrum disorders and not on maternal depression and anxiety [38][2]. This is a strong indicator that our work is one of the first of its kind for the MoBa dataset. Hence, an important aspect of this project is applying methods developed in the physical and computational sciences to learn how they perform in a new "environment".

Applying conventional statistical methods, such as multiple regression, to larger, more complex datasets are limited in their performance. Ultimately they can be poor at generalization and estimating non-linear effects [39], and can often be unstable when the number of variables is high due to multicollinearity [31]. The instability is related to the value of the coefficients, which can drastically change when minor adjustments are applied to the model. Utilizing machine learning and deep learning methods can remedy some of the obstacles associated with conventional methods. In this thesis, we will specifically apply elastic nets, neural networks and gradient boosted regression trees with the rich dataset from the Norwegian Mother, Father and Child Cohort Study to investigate the following research aims:

1. Investigate the ability of machine learning methods to predict a continuous measure of symptoms of anxiety and depression in new mothers, using data from a large population-representative prospective cohort.
    1.1. Predict levels of anxiety and depression at 6, 18 and 36 months postpartum using prenatal exposures measured at 17 and 30 weeks of pregnancy.
    1.2. Predict levels of anxiety and depression using exposures measured concurrently at 6, 18 and 36 months postpartum.
    1.3. Investigate the performance of modern machine learning approaches to statistical methods traditionally employed in psychology, such as linear regression.
    1.4. Compare the performance of models using i) aggregate scores on established scales, ii) item-level analyses, iii) dimensional reduction by principal component analysis, and iv) data without dimensional reduction
    1.5. Evaluate the performance of different methods for identifying individuals at risk for clinical levels of depression and anxiety.
2. Use variable importance methods to investigate whether machine learning methods can contribute to identifying prenatal or concurrent exposures that can be of clinical interest or help inform theoretical models of post-party emotional problems.
    2.1. Identify prenatal exposures associated with increased risk of symptoms of anxiety or depression at 6, 18 and 36 months after birth.
    2.2. Determine whether different sets of variables best predict internalizing symptoms at 6, 18 and 36 months in the extended postpartum period.
    2.3. Compare the sets of variables identified through traditional linear model methods and machine learning approaches.

[2]The article was published in January of 2022, highlighting the newly found interest in applying machine learning to the MoBa study.

3. Present a set of recommendations on using machine learning to analyze registry and health survey data based on the experiences from answering the first and second research aims.

The general capability to learn non-linear relationships, handle high dimensional data, and the overall versatility of the machine learning models and methods make their use compelling. We hypothesize that these abilities will lead to higher predictive abilities on a novel dataset. By focusing on prediction, machine learning could help identify individuals at risk for developing depression and could be a valuable tool in the screening process. At the same time, many women experience a decrease in several well-being measures following childbirth without developing any clinical diagnoses. Gaining insights into the non-clinical range of change is also of interest.

It is important to note that machine learning models are not without weaknesses. Often they are inappropriate for making statistical inferences. In theory, some methods can be used for both predictions and inference. However, the lack of an explicit model can make solutions obtained through machine learning challenging to relate to existing domain knowledge [40]. We are not blind to this shortcoming, but given the novelty of our work, we hope to show that, despite the inability to make statistical inferences properly, machine learning can provide new perspectives to a well-researched topic.

1.1 Terminology

As already demonstrated in the preceding section, the terms "traditional" and "conventional" methods will be used when referring to linear regression-based methods throughout this thesis, whereas "machine learning" is used to describe more complex, or sophisticated, statistical methods. We use the word "complex" to highlight several features associated with the models. It can refer to their structure, e.g., having deep connections through composite functions or a high number of parameters. Some machine learning models produce predictions that are difficult to explain, making it a complex task to unravel how they were produced. Other methods require an intricate implementation and are computationally complex. We note that, in all fairness, linear regression is a machine learning algorithm, more specifically, a supervised learning model. However, the context of use is important when referred to as a traditional method. As described above, linear models have traditionally been one of the most favoured tools for explaining statistical relationships. However, when used in a machine learning context, the focus is shifted from explaining to predicting. This distinction is important and is the reason behind our choice of phrasing in this thesis.

We introduced the term "deep learning", a subclass of neural networks with more than three layers. However, when this thesis mentions neural networks, we do not assume that the number of layers is less than three; they can have an arbitrary number of layers. So the two will be used interchangeably in our work. We give a formal introduction of the topic in chapter 6. Machine learning and deep learning are subfields of the larger research area called artificial intelligence.

Several terms related to the pregnancy and time period following childbirth are mentioned in the preceding section. We described the postpartum depression as depression that onset the first 12 months after giving birth, thus, implicitly defining the postpartum period as the first 12 months after delivery. The postpartum period does not have a strict definition regarding its endpoint in the literature. Some claim it lasts up to six months [41], while others use 12 months as the upper limit [7]. In our work, we will use the latter definition and work with what we coin the "extended postpartum period". This period describes the time from birth until the child is three years old. Prenatal is a term that is defined as the time the women are pregnant and the time leading up to giving birth.

1.2 Thesis Structure

This thesis is split into three parts: introduction and background, theory and method and results and discussion. A supplementary appendix is included, which contains a complete list of the raw data from the MoBa dataset used in our analyses and specific model configurations.

The first part of this thesis encompasses this introductory chapter, followed by a chapter covering necessary background material. Chapter 2 begins with a short description of the MoBa dataset and some related results, followed by a formal definition of depression and anxiety disorders and how they are assessed. The final part of the chapter includes a portrayal of the current state-of-the-art methods for longitudinal data, a discussion motivating the use of machine learning and the current state of affairs regarding the use of machine learning in health informatics before we present a discussion regarding prediction and explanation.

Theory and the methodological approach are outlined in Part 2. Here the first chapter covers some fundamental statistical concepts, while the following chapter, chapter 4, outlines a framework for supervised learning. The next three chapters give a detailed mathematical description of the numerical methods, starting with linear models in chapter 5, followed by neural networks and gradient boosted trees in chapter 6 and 7 respectively. The final chapter in Part II describes our methodological approach, giving a more thorough description of the MoBa study and the sample, the different measures included in the study and the procedures related to our statistical analyses. These procedures include how we processed the data to how the models were assessed and selected to meet the research aims stated above.

The third and final part of our thesis presents the results from our numerical experiments in chapter 9 and a discussion of the findings and their implications in chapter 10. Lastly, we conclude our work in chapter 11, where we summarize and give an outlook on future improvements.

Chapter 2

Background

This chapter gives a short introduction of the MoBa study (a more in-depth description is given in chapter 8) and previous studies related to the dataset. A more formal description of depression and anxiety disorders is given, followed by a section about known risk factors for the two. We include a section about how machine learning can enable us to take an agnostic approach when it comes to feature selection. This section gives an overview of conventional statistical methods associated with longitudinal data and what is considered the state-of-the-art method. Identifying the state-of-the-art method is motivated by wanting to align our choice of the conventional method with the literature. We describe the current state of machine learning in the healthcare domain, which includes studying electronic healthcare data and some of its challenges. Lastly, we describe the replication crisis currently taking place in several fields of science and how a shift in focus from explaining to prediction and incorporating some fundamental principles from machine learning may, to some extent, remedy this crisis.

2.1 Short Introduction of The Norwegian Mother, Father and Child Cohort Study

The Norwegian Mother, Father and Child Cohort study (Moba) is a study aimed to discover new knowledge about the causes of disease and health issues among mothers and children [42]. A cohort study is a type of longitudinal study design (see Appendix C for a description of longitudinal studies) that follows a participant sample over a specified time period. For the MoBa study, this period began at week 17 of pregnancy. When the participation period ended depended on when the mother was recruited. The study has, in some cases, collected data from the mother and child up till the child was 14 years old. The data was collected through self-reporting questionnaires.

The project was motivated by the lack of causal knowledge about diseases and their underlying factors, and it aimed to estimate the association between exposures and diseases for prevention reasons. The research design enables the researchers to identify healthy sub-cohorts within the participant group and monitor potential causal factors leading to the onset of a disease or health issue.

Due to privacy regulations, the data from the study is not publicly available, and all researchers have to apply to be granted access. The author has been given access through an ongoing project, meaning that the data available is mandated by this project. These regulations have practical implications for the thesis since the full scope of the data is not available in our numerical experiments. All the data related to the MoBa study is never given to any individual project.

2.1.1 Previous Studies on the MoBa Dataset

The use of Machine Learning

As stated in the introduction, as of March 2022, there have been 973 articles published based on MoBa. Out of these, there have been around 36 articles concerned with different aspects of depression and anxiety [37]. Among the publications, several direct their attention to parental or maternal symptoms of depression and/or anxiety and offspring consequences [43–46]. Others examine the prevalence of depression in smaller sub-cohorts, e.g., women with multiple sclerosis [18], or epilepsy [19].

To determine if any previous articles had applied any machine learning, we conducted several keyword searches to determine if any supervised learning models had been applied to the data. When we investigated the published articles' titles, none of the 973 articles mentioned any utilization of machine learning methods. Conduction the exact search in the search engine Google Scholar revealed that one article had applied a random forest and XGBoost model [38], while the use of the elastic net was found in less than a handful of articles. The keywords used in the search are listed in Table 2.1.

Table 2.1: An overview of keywords used when we searched the titles of published articles related to the MoBa study. Our search came back with no hits on the keywords listed below.

Keywords		
Machine Learning	Deep Learning	Tree-based models
Neural Networks	Artificial Neural Networks	Boosting
XGBoost	Elastic Net	Random Forest
Decision Trees	Support Vector Machines	Supervised Learning

The keywords in Table 2.1 were accompanied by the keywords "MoBa" and "Norwegian" when the search was conducted in the search engine.

Identifying Predictors of Depression and Anxiety

Data from MoBa has also contributed to identifying predictors of postpartum depression and anxiety. Sørbø et al. [20] assessed the association between adult physical, sexual and emotional abuse and postpartum depression. Associations between the different types of abuse and postpartum depression were performed with logistic regression. The reference group for the analysis was all the women who did not report any adult abuse. In total, 11% of the sample had postpartum depression, and 19% had experienced adult abuse. Women exposed to any type of abuse were 80% more likely to have symptoms of PPD compared to the reference group, while women experiencing all three types had an increased likelihood of 120%.

A recent study from late 2021 by Clayborne et al. [43] found that prenatal work stress was associated with both prenatal and postnatal depression and anxiety. Logistic regressions were conducted to examine the relationships between prenatal work stress and depression and anxiety. They found that experiencing higher levels of prenatal work stress is associated with increased odds of depression and anxiety both later in the pregnancy and the period following birth.

Body image has also been shown to mediate depressive effects associated with weight gain in the postpartum period for new mothers, and it is especially evident in obese women. Han, Brewis and Wutich tested the moderating and mediating effects of body image concerns on the emergence of new mothers' depressive symptoms [47]. The analysis was conducted with a binary logistic regression model with a discrete-time event history approach and a mediation analysis with bootstrapping. Experiencing a 10% increase in BMI in the postpartum period did not affect the likelihood of experiencing depressive symptoms for the first time for normal-

and overweight women. Women classified as obese had an increased likelihood of experiencing depressive symptoms by 18% in model 1. Normal- and overweight women with higher body image concerns were also more likely to experience symptoms of depression. The same was not found for obese women.

2.2 Depression and Anxiety

This thesis aims to predict levels of depression and anxiety amongst new mothers, which is a public health concern given the many negative consequences associated with it. A formal definition of the health conditions is given, followed by how they are typically assessed and known risk factors.

Definition and Prevalence

In a clinical environment, mental disorders are diagnosed using the framework established by primarily two classification systems: the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) and the International Classification of Diseases, Eleventh Revision (ICD-11). The description given of depression in ICD-11 is:

> Depressive disorders are characterized by depressive mood (e.g., sad, irritable, empty) or loss of pleasure accompanied by other cognitive, behavioural, or neurovegetative symptoms that significantly affect the individual's ability to function [48].

Depression is globally estimated to affect 5% of the general population [49], while postpartum depression affects around $10-15\%$ of new mothers. The corresponding definition of anxiety according is:

> Apprehensiveness or anticipation of future danger or misfortune accompanied by a feeling of worry, distress, or somatic symptoms of tension. The focus of anticipated danger may be internal or external.[48].

A review study estimated the global prevalence of anxiety disorders to be around 7% [50]. As described in the introduction, between $15-20\%$ of new mothers experience postpartum anxiety.

Assessment

When diagnosing depressive and anxiety disorders, both in the postpartum period and in general, individual clinical interviews are considered to be the gold standard [51]. A clinical interview is a tool designed for mental health professionals which involves gathering information about the patient through observations and dialogue. In these interviews, severity and symptoms are assessed using clinician-rating scales instead of self-reporting scales. Unfortunately, there is no equivalence between outcomes of the clinician-rated and self-reporting scales; they both provide unique information that needs to be considered when making a diagnosis [52]. Using these clinician-rating scales is a time-consuming and costly practice, explaining why self-report instruments have a long tradition in psychotherapy research. Self-report questionnaires return a score used to estimate the respondent's likelihood of showing clinical levels of disease using established cut-off thresholds. Several studies have shown that the use of self-reporting scales leads to an overestimation of prevalence, especially in true low-prevalence populations [53].

One such self-reporting questionnaire is developed to measure symptoms of several mental disorders, two of them being anxiety and depression, the Hopkins Symptom Checklist (SCL) [54]. It is a 90 items checklist; here "item" refers to a question in the checklist. Several short versions of the original instrument have been created. Among them are a 25-item version (SCL-25), a 5-item version (SCL-5) and an 8-item version (SCL-8). All items are scored using a Likert

scale [55] from 1 to 4, with 4 being the highest indicator of depressive or anxiety symptoms. For all versions of the symptom checklist, the total SCL score is the mean of all responses. A mean score greater or equal to 1.75 is considered to be a strong predictor of a mental disorder in need of treatment in the SCL-25 [56]. Several cut-off points have been established for the Norwegian population. For SCL-10, the cut-off is 1.85, while for SCL-5, it is 2.0 [57].

A symptom scale is explicitly designed to detect postpartum depression: The Edinburgh Postnatal Depression Scale (EPDS). The original scale consists of ten short statements, and a short version comprised of five statements, EPDS-5, has been developed and validated in a Norwegian population-based sample of pregnant women. The short-scale correlates 0.96 with the original EPDS, and 0.75 with SCL-25 [58].

Risk Factors

The literature documenting the consequences of postpartum depression and anxiety is rich, and they have been associated with a significantly increased risk of difficulties in the relationship between the mother and child and of problems in the child's attachments in general [59]. Adverse effects on the relationship between the parents with maternal postpartum depression have been found to be one of the strongest predictors for paternal postpartum depression [60] and an increased risk for future maternal depressive episodes [6]. The literature demonstrates that depression and anxiety experienced after childbirth is a problem that affects the entire family and has far-reaching consequences in other areas of life.

Studies aimed at identifying risk factors for postpartum depression and anxiety are usually cohort studies, where the women are recruited during pregnancy and followed throughout the postpartum period [20, 61, 62]. Several risk factors for the two have been identified through such studies, with some of the strongest predictors being a family history of psychopathology, history of depression, psychological disturbance during pregnancy, lack of social support, poor marital relationship and socioeconomic adversity [5, 10, 20, 63]. Immigration status is often linked to socioeconomic status, which could also be a risk factor when the sample is of multicultural background. Immigrant women also tend to have a higher chance of developing depressive symptoms in the time after delivery compared to non-immigrant women [64]. It has been found that some of the predictors are culturally dependent, with cultures being more rooted in traditional household norms exhibiting risk factors connected to these norms, e.g., husband's unemployment and pressure related to the baby's sex [63]. Postpartum- and labour complications are also associated with an increased risk of developing postpartum depression and anxiety [65, 66].

Considering individual personality characteristics is important when determining risk factors for mental health disorders such as depression and anxiety. Personality influences how people perceive and respond to their surroundings, making it a determinant of behavior. Neuroticism, also known as negative affectivity (NA), has been linked to an increased risk of mood disorders like depression [67]. It is a broad personality trait that reflects pervasive individual differences in negative emotionality and self-concept; it describes the tendency to experience negative emotions frequently and to have a negative view of oneself [68]. Self-efficacy, a personality trait reflecting an individual's belief in his or her capacity to execute behaviors necessary to produce specific performance attainments [69], also plays an essential role in models designed to explain health behaviour. High levels of self-efficacy are associated with effortful engagement in tasks, willingness to change undesired behaviour and effective sustainment of changed behaviour [67].

2.3 Taking An Agnostic Approach with Machine Learning

Conducting statistical analyses in the field of psychology, using what we coin as conventional methods, often involves a modest number of independent variables [27]. The independent variables included are often deemed as confounding variables, meaning that they influence both the independent- and dependent variables and can be identified using domain knowledge and theory [70]. There are several reasons behind using a limited number of features, e.g., it can make the results more interpretable and circumvent possible methodological limitations related to high dimensional data. However, selecting the amount and which features to include correctly can be a challenging task.

By introducing machine learning, we can employ numerical methods more suited for handling high dimensional data and thus expanding the feature space. Therefore, we can take a more agnostic approach when it comes to feature selection. Here, agnostic refers to not making any assumptions about which method, if any, is best suited for our specific purpose. A part of the first research question is to identify different scenarios where the different models exhibit high predictive performance. More specifically, these scenarios will be different datasets with a varying number of features. This will allow us to investigate how the predictive ability will be affected as a function of the number of features.

2.3.1 State-of-The-Art Longitudinal Studies

Here we present an overview of several conventional statistical methods used when working with longitudinal data, along with an attempt at trying to establish which of these are considered to be the state-of-the-art method. The state-of-the-art model will represent the conventional method in the field of psychology later in the thesis. Some of the techniques described in this section will be elaborated upon in the second part of the thesis, more specifically in chapter 5. Others will receive a more high-level treatment as they are not imperative to the rest of the thesis. We preface the section by saying that there is not necessarily *one* method that turns out to be a single victor in this search. Oftentimes, the analysis method can depend on several factors, e.g., the type of data and its complexity, theoretical background, and thus the choice of model. However, it is still possible to ascertain which methods dominate the field by examining the literature.

A paper reviewed several statistical techniques for analyzing longitudinal data for neurodegenerative diseases, using examples from Huntington's disease studies [71]. The statistical techniques featured in the article were Change Score Analysis, (M)ANOVA approaches, Generalized Estimating Equations (GEE) and Mixed Effects Regression (MER). Before discussing its findings, we give a brief description of the methods.

Change Score Analysis analyzes the difference between measurements at each time point, with the change score being the difference in the dependent variable at two time points. Change scores can be used as the dependent variable in a regression analysis.

(M)ANOVA, being short for (Multivariate) Analysis Of Variance, is a statistical tool for detecting differences between experimental group means. It applies to designs with one continuous parametric numerical dependent variable and one or more independent variables [72]. The use of (M)ANOVA requires a complete dataset, i. e. no missing data and that all participants are measured at the same number of time points.

Two more modern techniques are GEE and MER. They can handle dependent variables that are either time-invariant or time-varying, missing data and irregularly spaced measurements with respect to time. They can also model time-dependent predictors.

Generalized Estimating Equations is designed for analyzing the regression relationship between covariates and repeated measurements. A correlation structure, or model, which describes the correlation between the repeated measurements is defined prior to the analysis,

meaning that a possible incorrect model may be chosen. Hence, the use of GEE is not suitable if the objective is to obtain information about the correlation structure [71].

Mixed Effect Regression, whose theory will be developed in chapter 5, on the other hand, provides information on the regression relationship between covariates and repeated measurements and the correlation structure. The MER framework is an extension of multiple linear models, a theoretical background is included in Appendix A, and incorporates so-called fixed and random effects. Fixed effects are variables that are constant across individuals and estimate the average population trajectory, while random effects vary and capture the variation of the individual trajectories around this average [73, 74]. The random effects also capture the correlation between repeated measurements. We can thus describe the difference between the two in the following manner: The fixed effects correspond to the independent variables from the multiple linear regression case, while the random effects are usually considered grouping factors used to control the observed variance in the response variable.

The paper found the MER framework to be the most flexible when dealing with longitudinal data and its related challenges [71]. It can handle missing data without the need for imputation, and it can also provide subject-specific estimates. In the field of modeling neurodegenerative diseases, MER models have become the standard for modeling correlated longitudinal data [71].

Another family of methods widely used when investigating development over time is latent-curve modeling (LCM), which is a part of a bigger framework, namely structural equation modeling (SEM). The primary focus of SEM is on testing causal relationships anchored in theory [75] and includes several different methods. The underlying concept of LCM is identical to MER's: An overall mean trajectory is estimated for the entire sample size, while individual differences are encoded in the random effects. In LCM, random effects are specified as latent variables, those being variables that are not directly observed but rather inferred from other variables. A paper by McNeish and Matta [76] compared the two modeling approaches on longitudinal data and found that MER is better suited for complex data structures and straightforward models, while the opposite is said to be true for LCM. In general, the implementation of SEM methods is more complex than MER models, and when mediating variables are not present in the dataset, MER models exhibit a good performance with respect to statistical power relative to SEM [77].

A meta-study of current practices in data analysis methods in psychology from December 2018 found that papers published in the journal *Health Psychology* were more likely to either involve analysis of relationships among variables using regression models or SEM, with a prevalence of 51.85% and 22.22%, respectively [28]. Based on said flexibility, implementation and prevalence in research, linear regression models will serve as the models of comparison against the machine learning models.

2.3.2 Machine Learning in Health Informatics

Health informatics is an interdisciplinary field of science that aims to develop and assess methods and technologies for organizing, acquiring and analyzing patient data. It encompasses fields like medical imaging, bioinformatics, medical informatics and artificial intelligence in healthcare [78]. In the age of big data and an increasing amount of patient data stored electronically, the use of machine learning (ML) methods in the field of healthcare have had a rapid increase in the last decade [79]. Due to varying availability and privacy regulations for different types of patient data, some subfields, such as medical image processing, have experienced a higher degree of advances and progress in implementing ML methods compared to, e.g., longitudinal electronic healthcare data.

Progress and Benchmarking

Assessing and discussing the progress of ML methods in any field of research requires benchmark datasets to be established as well as meaningful evaluation metrics. This makes it possible to directly compare the performance of separately developed models on a prediction task. Given the sensitive nature of patient data, there is a lack of benchmark datasets, and it is considered to be a serious impediment to the advances of ML for electronic healthcare data [80]. The absence of publicly available datasets reduces the reproducibility of scientific findings, given that many studies use private datasets. This thesis is no exception to this trend. It also makes it challenging to establish state-of-the-art methods for this type of data, given that they can not be compared and verified by outside parties [81].

The lack of publicly available benchmarking datasets is not the same as a total absence of them. The Medical Information Mart for Intensive Care (MIMIC) dataset[1] is one such dataset that is used for direct comparison of models designed for prediction tasks on longitudinal health data. However, when comparing the predictive performances of ML methods on mortality based on the MIMIC dataset and their progress over the last five years, the results are disheartening: The progress has been relatively stagnant, and deep learning models do not outperform the traditional regression approaches [80]. These results can be somewhat surprising, given the ML methods abilities to model complex relationships. The findings raise questions about the choice of dataset, evaluation metrics and the preprocessing of the data and model choices. There could also be that the deep learning models are not more suitable for these types of prediction tasks.

Interpretability and Explainability

As mentioned in the preceding section, linear models dominate the field of health psychology, and the same is said to be true for the field of applied clinical informatics. One of the most important reasons behind this is the interpretability of linear regression models. In the field of machine learning, interpretability refers to which degree the produced outcome from a model is predictable or understandable to human beings [82]. For clinical applications of ML models, the predictions and how they are made are paramount for medical health professionals. They may have real-life consequences on medical decision-making or treatments. Many ML techniques are described as "black boxes" when making predictions. The black box analogy refers to the input and output being the only parts of the method that convey meaning to a human observer [81]. Methods deemed more interpretable are often coined "white boxes". Hence, the area of research that focuses on deep interpretable models for electronic healthcare data is active and expected to be an ongoing research area for the years to come.

We note that in discussions about the interpretability of ML models, the term explainability is often mentioned within the same context. There is a subtle distinction between the two, but they are often used interchangeably in some literature [83]. Explainability is associated with the inner mechanisms that take place when a model is making predictions and how humans can achieve an understanding of these procedures [84]. Due to the inconsistent use of the two terms, some researchers have proposed new terminology to avoid ambiguous use, such as opaque systems to describe models that offer no insights into how the predictions are made [85].

So-called white box methods are more explainable and interpretable than the black box models. However, they often show lower predictive abilities and fail to achieve state-of-the-art performance when compared to the more sophisticated models [84]. The development of trustworthy interpretation algorithms, algorithms that explains how deep models make decisions, is vital in order to be able to apply more sophisticated models to sensitive and high-stakes decision tasks that are often found in the health care sector. In the European Union, so-called automated individual decision-making algorithms are subject to the General Data Protection Regulation. This regulation gives the individual right to receive a full explanation behind an

[1] https://mimic.mit.edu/

algorithmic decision [86]. So in the future, if machine learning models contribute to screening processes or make decisions regarding treatments, there is a need for explainable models.

2.3.3 To Explain or Predict?

There is an ongoing discussion in the field of statistics about how statistical modeling should best be exercised, and machine learning is at the heart of this conversation. The discussion is centered around statistical modeling for causal explanations and predictions and which of the two is the most suitable framework to work within. It has been ongoing for the last 20 years and was fueled by the article "Statistical Modeling: The Two Cultures" [87] by Leo Breiman from 2001.

The article taxonomizes statistical models based on their structure and how this structure enables the models to produce reliable information about the underlying data structures. He presents two cultures of modeling: data- and algorithmic modeling, where the former produces outputs through parametric models and the latter with non-parametric models. He coins the non-parametric models "black boxes". In relation to our work, the conventional methods belong to the data modeling culture, while the two more sophisticated machine learning methods of interest, neural nets and boosting, belong to the algorithmic modeling culture. Breiman argued that the statistical community almost exclusively used data models and that this had led to irrelevant theory development and questionable scientific conclusions [88]. He further stated that the assumption that complex data emerging in the natural sciences assumed to be generated from an a priori parametric model could lead to problematic conclusions. Describing observations through such a model imposed a straight jacket on the ability of statisticians to deal with a high number of statistical problems. The proposed solution was to shift focus to algorithmic modeling, which has predictive accuracy as its primary goal. His article touches upon the discussion from the preceding section. He argues that the overall goal of any model is to obtain useful information about the relationship between the dependent and independent variables, which is not the same as interpretability.

Breiman did not argue for a Copernican shift regarding how statisticians should conduct their work, but rather that it is essential to keep an open mind when faced with a new problem. The best solution may be a data model, an algorithmic model or a combination of the two. The article met mixed responses, ranging from a more or less full agreement, having some reservations, while others completely disagreed[2]. A paper by Shmueli in 2010 [89] expanded Breiman's point about the two modeling cultures and argued that in many scientific fields, statistical models are used almost exclusively for causal explanation, as opposed to empirical prediction. The models used for causal explanations are what Breiman referred to as data models, while algorithmic modeling was suited for predictions. In the social sciences, prediction is often considered unscientific, but both explanatory- and predictive modeling is needed for generating and testing theories [89]. Shmueli argues the two approaches should not be viewed as opponents, but an increased focus on prediction can be viewed as a complementary goal that can ultimately increase the fundamental theoretical understanding [89]. The final statement speaks to what is considered one of the goals of this thesis.

The Replication Crisis

In many fields of science, from psychology to physics education research (PER) to medicine, there is an ongoing replication crisis where journals publish statistically significant results that do not hold up when the same experiments and analyses are independently conducted at a later date [25, 90, 91]. There is some evidence that points to this crisis being partially a result of studies having a near-exclusive use of statistical modeling for causal explanation, together with

[2]Commentaries from four known statisticians are attached in [87].

the assumption that models with high explanatory power automatically have high predictive power and the use of small sample sizes [89, 92, 93]. There is also a tendency among researchers, reviewers and editors to favor results that are deemed as "good", meaning that they are either more worthy of publication or more in agreement with an underlying hypothesis, which can lead to "p-hacking" [94, 95]. P-hacking is the procedure of selectively choosing analytical procedures based partly on the quality of the results they produce (p-values and their use in hypothesis testing are described in Appendix A). This can lead models with high explanatory power to fail to predict similar behaviour in novel data due to being overly specialized to the data at hand. This is known as overfitting. Without disregarding the scientific value of explanatory modeling, several researchers argue that incorporating selected core principles from machine learning, such as predictive modeling and the use of cross-validation to avoid overfitting (see chapter 4), can, to some extent, mitigate this crisis [27, 92]. The rest of this discussion centers around the field of psychology but also applies to other scientific fields.

Psychology has traditionally been concerned with explaining the underlying mechanisms that govern observable behaviour and mental processes. This has led the field to have a natural inclination to emphasize a model's explanatory power and its ability to provide theories about psychological mechanisms instead of favoring high predictive ability. This can, in the worst-case scenario, contribute to the ongoing replication crisis due to models being overfitted [27]. Statistical models are often evaluated based on how well they fit observed data, but the goodness of fit does not guarantee high prediction accuracy on unseen data. So the assumption that high explanatory power equals high predictive power is not necessarily true from a statistical standpoint. To understand this, we make a small introduction of what is known as the bias-variance tradeoff in machine learning, which is discussed later in chapter 4.

In short, the bias-variance tradeoff makes it impossible for a statistical model to have low variance and a low bias, which is the desirable outcome. They are both terms in the expected generalization error a model makes when using a quadratic loss, together with the irreducible error term (see Appendix B for a derivation). The variance is a measure of dispersion, and bias is the error made due to an erroneous statistical model. Achieving low variance requires a low number of parameters; alas, this would lead to a high bias. Hence, there is a conflict that needs to be resolved with a tradeoff. When aiming to fit a model to observed data, good results are obtained by minimizing the bias, leaving the variance unattended. This can lead to practically useless models for prediction, which again can cause failure in replicating the results in the future. By contrast, in machine learning, the fundamental goal is to construct a model that predicts well on unseen data, which is achieved when the total prediction error is minimized.

Another issue related to explanatory modeling in psychology is the sample size and "extraordinary" discoveries. When the sample size is small, it becomes easier to obtain larger effect sizes due to more variable estimates. Combined with the trend of favoring "interesting" results, these findings would end up being published, and this explains the high number of remarkable discoveries reported in the mid-1990s in the field of quantitative genetics [27]. Had the researchers taken a predictive approach, there is a chance that the final results would have been less extraordinary, given that the original participant sample may not be representative of the general population. In the era of big data, the number of extraordinary discoveries has declined, and big data studies tend to report more modest effect sizes. This decline partially lies in the studies' sample sizes and predictive nature, together with that previously reported effect sizes in many domains were never truly big to begin with [89]. The argument is not to abandon explanatory modeling for good but rather emphasize that an increased focus on prediction can lead to a greater understanding of behaviour.

Part II

Theory & Method

Chapter 3

Statistical Concepts

Before venturing into the domain of machine learning, we need to take a detour and cover some statistical concepts that are crucial for understanding the chapters to come. More specifically, we will describe the bivariate normal distribution, a special case of the multivariate normal distribution and maximum likelihood estimation.

3.1 The Bivariate Normal Distribution

The bivariate normal distribution is a special case of the multivariate normal distribution. An $N \times 1$ random vector $\boldsymbol{x}$ is N-variate normally distributed if every linear combination of its components are univariate normally distributed, which can be summarized with the notation $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, V)$. Here $\boldsymbol{\mu} \in \mathbb{R}^N$ is the mean vector, and $V \in \mathbb{R}^{N \times N}$ is the covariance matrix, where the element at the ith row and jth column is

$$V_{ij} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] = \mathrm{Cov}[x_i, x_j], \tag{3.1}$$

for $i, j = 1, ..., N$. The variance vector is thus the diagonal of V, given that $\mathrm{Var}(x_i) = \mathrm{Cov}[x_i, x_i]$. The joint density function for the random vector $\boldsymbol{x}$ is expressed as

$$f(\boldsymbol{x}) = (2\pi)^{-n/2}(\det V)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T V^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{3.2}$$

We observe that the density in eq. (3.2) reduces to the univariate normal distribution when $\boldsymbol{x}$, $\boldsymbol{\mu}$, $\mathrm{Diag}(V) \in \mathbb{R}^{1\times 1}$. In the section about linear mixed models (section A.2) we encounter a bivariate normal distribution, hence we restrict the rest of this section to the special case where $\boldsymbol{x} \in \mathbb{R}^2$.

For a two-dimensional vector, the notation $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, V)$ translates to

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_2, x_1) & \sigma_2^2 \end{pmatrix}\right), \tag{3.3}$$

where σ_i^2 is the variance of variable x_i for $i = 1, 2$. It can be shown that the conditional expectation of x_1 given x_2 is [96]

$$\mathbb{E}(x_1|x_2) = \mu_1 + \mathrm{Cov}(x_1, x_2)\sigma_2^{-2}(x_2 - \mu_2). \tag{3.4}$$

3.2 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method for estimating the parameters $\boldsymbol{\theta}$ which specifies a probability function $p(X|\boldsymbol{\theta})$. The function describes the probability of observing a discrete, or continuous, variable X for a given number of unknown parameters $\boldsymbol{\theta}$ and is called

a likelihood function. The likelihood function is a function of the parameters $\boldsymbol{\theta}$, with X being held fixed. This leads to an important distinction: as long as the parameters are unknown, the likelihood function is not a probability mass- or density function and is often written as $l(\boldsymbol{\theta}|X)$.

We assume the observations of X to be independent of each other, hence $l(\boldsymbol{\theta}|X)$ can be expressed as the product

$$l(\boldsymbol{\theta}|X) = \prod_{i=1}^{N} p(X = x_i|\boldsymbol{\theta}), \tag{3.5}$$

with $x_1, ..., x_N$ being the observations in X. The principle in MLE is to choose the parameters that maximizes the likelihood of observing the data X, more formally expressed as the following optimization problem

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|X). \tag{3.6}$$

Given that the likelihood function is a product, one can equally estimate the log-likelihood to simplify calculations (since the logarithm is a monotonic function).

3.2.1 Profile Likelihood

Sometimes we are not equally interested in all of the parameters in $\boldsymbol{\theta}$. Suppose $\boldsymbol{\theta}$ can be partitioned into $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, with $\boldsymbol{\psi}$ being the parameters of interest, and $\boldsymbol{\lambda}$ being what is called nuisance- or hyperparameters. A nuisance parameter is any parameter that is not of direct inferential interest. Even though the hyperparameters are not of direct interest, they have to be estimated along side $\boldsymbol{\psi}$. To do so, we assume $\boldsymbol{\psi}$ is known, thus making the likelihood function only a function of $\boldsymbol{\lambda}$,

$$l(\boldsymbol{\psi}, \boldsymbol{\lambda}) = l_{\psi}(\boldsymbol{\lambda}). \tag{3.7}$$

The notation l_{ψ} indicates that $\boldsymbol{\psi}$ is being held fixed. The estimated nuisance parameters, $\hat{\boldsymbol{\lambda}}_{\psi}$, are then obtained from maximizing eq. (3.7) with respect to $\boldsymbol{\lambda}$. Substituting $\boldsymbol{\lambda}$ with $\hat{\boldsymbol{\lambda}}_{\psi}$ into the original likelihood function results in the profile likelihood function $l_p(\hat{\boldsymbol{\lambda}}_{\psi}, \boldsymbol{\psi})$. The profile likelihood is thus a joint likelihood where the nuisance parameters are expressed as functions of the parameters of interest, i. e. $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$.

For each $\boldsymbol{\psi}$ we will have a new curve $l_{\psi}(\boldsymbol{\lambda})$. To estimate $\boldsymbol{\psi}$, we evaluate the maximum $l_{\psi}(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda}$, and choose the $\boldsymbol{\psi}$ that yields the highest value of $l_{\psi}(\boldsymbol{\lambda})$. This translates to

$$\begin{aligned} \hat{\boldsymbol{\psi}} &= \arg\max_{\psi} l_{\psi}(\hat{\boldsymbol{\lambda}}_{\psi}), \\ &= \arg\max_{\psi} l\big(\arg\max_{\lambda} l_{\psi}(\boldsymbol{\lambda}), \boldsymbol{\psi}\big). \end{aligned} \tag{3.8}$$

Chapter 4

Supervised Learning Framework

In this thesis, we explore supervised learning methods on longitudinal healthcare data for prediction purposes. Supervised learning refers to machine learning algorithms that fit a function that maps an input to a known output based on observed data. Throughout this chapter, we will use the word function and model interchangeably. There are many different supervised learning algorithms, but some concepts are fundamental for all of them. This chapter outlines the framework that the vast majority of supervised learning algorithms work within.

Supervised learning typically deals with two types of problems, regression and classification. The difference lies in their predictions; regression algorithms predict continuous values while classification algorithms predict discrete outcomes. The following sections will focus on supervised learning algorithms used to solve regression problems, given the nature of this thesis.

We have organized the chapter into five sections. Each section can be interpreted as a different step in an analysis pipeline. We do not try to present a complete workflow that applies to every supervised learning problem but rather an overview of some essential aspects in need of consideration when faced with a new problem. There exists a vast amount of research on all of the different sections in this chapter. Hence, it can be thought of as merely a summary of some selected literature from the field.

We start the chapter by describing the underlying objective behind every supervised learning algorithm before diving into how to handle and process raw data. In section 4.3 we describe how a supervised learning algorithm can be assessed before the process of model selection is presented. The final section, section 4.5, is about model optimization, and we present three different optimization algorithms.

4.1 Objective

Suppose we have a dataset on the form $\mathcal{D}(X, \boldsymbol{y}) = \{(\boldsymbol{x}^{(i)}, y_i)\}_{i=1}^{N}$ where N is the number of observations, $\boldsymbol{x}^{(i)} = (x_{i1}, \ldots, x_{ip})$ is an $1 \times p$ feature vector for each observation i and y_i is the ith dependent variable in the response vector $\boldsymbol{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$. Note that the formalism presented can easily be extended to the case where the response variable is multivariate. The feature vectors can be placed in an $N \times p$ matrix X, with $\boldsymbol{x}^{(i)}$ as the the ith row vector. The observed data is assumed to be generated from an unknown function f_{true}.

The goal is to explain y_i through a mapping function $\hat{f}(\boldsymbol{x}^{(i)})$, where $\hat{f}$ is chosen from what is called a hypothesis set $\mathcal{H}$. The hypothesis set consists of all functions we are willing to consider as a plausible explanation for our observed data. When choosing a supervised learning model to fit the data, assumptions about the hypothesis set and mapping function are made. The different functions, or models, can be divided into two sub-groups: parametric and non-parametric models.

Parametric Models These models assumes that the observed data can be explained through a mapping function with a fixed number of parameters $\boldsymbol{\theta}$, meaning that they are independent of the observed data. The true underlying function is assumed to be a noisy model and a function of the unknown true parameters $\boldsymbol{\theta}_{\text{true}}$,

$$y_i = f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}_{\text{true}}) + \varepsilon_i \qquad \text{for } i = 1, ...N, \tag{4.1}$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ being independent and identically distributed (i.i.d) noise with zero mean and standard deviation σ. Each observation has an inherent irreducible error, making $\boldsymbol{\varepsilon}$ a vector of length N.

Non-Parametric Models Non-parametric models, on the other hand, do not make strong assumptions about the mapping function and its parameters. The structure of $\hat{f}$ is instead determined by the observed data $\mathcal{D}$. The name "non-parametric" does not refer to the model having zero parameters; it simply means that the number of parameters is adjustable. Given that the nature and number of parameters are not assumed to be known a priori, the true underlying function for a non-parametric model has the same form as eq. (4.1), except that the true parameters are now some optimal parameters $\boldsymbol{\theta}_{\text{opt}}$.

Fitting the mapping function $\hat{f}$ to the data $\mathcal{D}$ is done through estimating the parameters $\boldsymbol{\theta}$ in a way that minimizes a defined error function, also called a loss or cost function, that quantifies the deviation of the predicted outcome from the true response. In mathematical terms this translates to

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{C}(\boldsymbol{y}, \hat{f}(X; \boldsymbol{\theta})), \tag{4.2}$$

with $\mathcal{C}(\boldsymbol{y}, \hat{f}(X; \boldsymbol{\theta}))$ being the cost function. Here the notation $\hat{f}(X; \boldsymbol{\theta})$ refers to $\hat{f}$ being evaluated and summed over the matrix X row wise. When estimating the parameters that best fit the data, we say that we are training the model. If $\hat{f}$ is approximately equal to f_{true}, the model is said to generalize well, meaning that we would be able to make meaningful predictions on unseen data. After the training is complete, $\hat{f}$ is then a regressor (or classifier in the case where we are predicting discrete outcomes).

We add a small note about the minimization problem in eq. (4.2): it can also be recast into a maximization problem if we describe our observations through a likelihood function as described in section 3.2. Throughout the theory part of the thesis, we will mainly be framing our model fitting process as a minimization problem.

4.2 Handling Data

After choosing an appropriate mapping function for our problem, the first step is to prepare our data for modeling. This step is of the highest importance, and how well the data is prepared can significantly impact the final result. The process of preparing the raw data that goes into the learning algorithm is referred to as preprocessing. The algorithm determines how the data is processed, as different algorithms require different data types. In general, we categorize data into four different data types:

1. *Nominal Data*: This data type has no quantitative value, meaningful ordering or well-defined zero. This form of data is often used for labeling. An example of a nominal variable is nationality, which can take several values, but has no ordering or meaningful zero.

2. *Ordinal Data*: Here, the data is categorical with a meaningful ordering, but the distance between values is not equal. As with nominal data, ordinal data has no well-defined point of zero. This type of data is often used to measure opinions using a Likert scale. The SCL questions in Table 8.1 are examples of ordinal variables, where the different answers are integer encoded.

3. *Interval Data*: This type of data is measured along a numerical scale which gives the data points a meaningful ordering, and the distance between values is now equal. This data type has no true zero but could have an arbitrary point treated as zero. Measurements of temperature with either Celsius or Fahrenheit as scales are examples of interval data.

4. *Ratio Data*: The final data type has all of the same properties as interval data, except there exists a well-defined zero. Hence, height, age and weight measurements are all examples of ratio data. Ratio data are always greater or equal to zero, i. e., never negative.

This classification of data types was introduced in 1946 by the psychologist S. S. Stevens [97], and the two first categories are usually described as qualitative data types, while the remaining two are quantitative.

Depending on the choice of learning algorithm and the nature of the data $\mathcal{D}$, it might be necessary to transform parts of data before being able to perform the desired analysis. Many real-life datasets contain a high number of measured features and they are more likely than not to contain missing values and outliers. These problems should be addressed before beginning the training procedure and are covered later in the chapter.

4.2.1 Splitting Data

To evaluate the predictive performance of the model on unseen data, we usually split the dataset $\mathcal{D}$ into two partitions: a training set $\mathcal{D}_{\text{train}}$ and a test set $\mathcal{D}_{\text{test}}$. We assume that we are in a data-rich situation, meaning we have sufficient data in each split. The ratio for splitting the dataset can vary, as there is no general rule on choosing the partition of data points in each set.

The training data can again be partitioned into a validation set, $\mathcal{D}_{\text{val}}$, with the remaining data being used as training data. The validation set gives an unbiased evaluation of the results from the training set and can be used for determining hyperparameters. If data is scarce and there is not enough data for this three-fold split of the data, resampling techniques can be applied. In subsection 4.4.4 we introduce one such technique, namely k-fold cross-validation. After splitting the data, the different sets are shuffled randomly to eliminate any human bias concerning the ordering of data points.

When working with data in the long format, i. e. each observation unit has multiple rows in the data matrix X; data from a unique unit must appear exclusively in one of the sets.

4.2.2 Transforming Data

One-Hot-Encoding

Ordinal data have a natural order, making it easy to encode with integers. On the other hand, nominal data do not have an inherent order, making their encoding trickier. One way to remedy the missing order in nominal data is by one-hot-encoding, which can also be applied to ordinal data.

The concept of one-hot-encoding is that when dealing with a variable with M categories, we introduce M new binary variables for each unique category and remove the original variable. The correct category variable is given the numeric value one, while the remaining categories are set to zero. The two tables 4.1 and 4.2 provide an example of how nominal data can either be encoded using labels (Table 4.1) or one-hot encoded (Table 4.2), with the nominal variable holding the answer to the question "Which fantasy franchise do you prefer?", with the following answer alternatives: the Lord of the Rings (LOTR), Game of Thrones (GOT) or Harry Potter (HP).

Table 4.1: An example of nominal data being encoded with labels

Person ID	Preferred Franchise
1	GOT
2	LOTR
3	GOT

Table 4.2: An example of how label encoded data can be one-hot encoded, based on the data in Table 4.1.

Person ID	LOTR	GOT	HP
1	0	1	0
2	1	0	0
3	0	1	0

Feature scaling

Raw data can widely vary, and machine learning algorithms may not work as intended without proper scaling. Two very common ways to transform the data into having the same scale is through standardization or normalization [98].

Standardization When standardizing the data you subtract the mean value of the dataset (not the complete set $\mathcal{D}$, but one of the subsets) and then divide it by the standard deviation. The data will then have mean zero and a standard deviation of one. For an arbitrary vector $\boldsymbol{v} = (v_1, \ldots, v_m) \in \mathbb{R}^m$, the scaled vector element, v_i', is

$$v_i' = \frac{v_i - \mathbb{E}(\boldsymbol{v})}{\sigma_v} \qquad \text{for } i = 1, ..., m, \tag{4.3}$$

with $\mathbb{E}(\boldsymbol{v})$ and σ_v being the expectation value and standard deviation of $\boldsymbol{v}$, respectively. This scaling would the be applied to every column-vector in the design matrix X.

Normalization When we normalize our data we transform it so that every value lies in the range $[0, 1]$. For each value we subtract the minimum value, and then divide it by the maximum value subtracted by the minimum. This scaling method is also referred to as min-max scale. For an arbitrary vector $\boldsymbol{v} = (v_1, \ldots, v_m) \in \mathbb{R}^m$, the scaled vector element, v_i', is

$$v_i' = \frac{v_i - \min(\boldsymbol{v})}{\max(\boldsymbol{v}) - \min(\boldsymbol{v})} \qquad \text{for } i = 1, ..., m, \tag{4.4}$$

Normalization can be sensitive to outliers, whereas standardization is not. All of the numerical experiments in this thesis will thus be performed on standardized data.

Regardless of the scaling procedure, the scaling should be applied separately for the data partitions $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ to prevent data leakage between the train and test set. Data leakage occurs when information from the test set becomes known to the training set and affects the estimation procedure, thus making the test dataset no longer completely unseen. It can lead to an overly optimistic out-of-sample error that will not reflect the model's true performance on unseen data.

4.2.3 Dealing with Outliers

Outliers are data points that lie far from the mean of the observed data, and including them in the fitting procedure can affect the model's predictive ability. If the data is normally distributed, it is customary to define a threshold of 1, 2 or 3 times the standard deviation to define outliers [99, p. 107]. Observed data above this threshold are removed. Outliers can arise from a noisy experiment, an experimental error, human errors or data variability. If the three former reasons apply, the observations are best removed, while if it is the latter case, the data points should be kept.

4.2.4 Imputation

Imputing a dataset is one way to tackle missing data. Wrongful treatment of this type of data can affect the statistical power of an analysis or lead to misleading conclusions. There are a collection of imputation techniques with varying complexity, and choosing the right one for a particular dataset is not necessarily a clear cut. We restrict this section to briefly describe the different missing-data mechanisms and describe a small selection of imputation techniques. We stress that the data partitions should be imputed separately.

Missing-data Mechanisms

Knowing the underlying mechanism to why the data is missing can be helpful when deciding the imputation method. Rubin [100] categorized these into three groups:

1. *Missing Completely at Random (MCAR)*: If the probability of an observation being missing is uniform for all independent variables, the data point is MCAR.

2. *Missing at Random (MAR)*: A data point is MAR if the probability of it being missing only depends on the information available in the dataset.

3. *Missing Not at Random (MNAR)*: When data is MNAR, the probability of it being missing varies, and the reason behind these fluctuations is not explained through the data.

In most cases, the MCAR assumption is unrealistic and overly simplified. The MAR assumption becomes more feasible by including as many predictors as possible. In an ideal world, all predictors that affect the probability of the data being missing should be included if the data is MAR. Wrong assumptions about the mechanisms above can lead to biased estimates. We assume that the missing data in the MoBa dataset is MAR.

Imputation Techniques

Below is a short description of some selected imputation techniques; for a more complete overview, we refer the reader to [101, 102].

Removing Data Many imputation strategies solve the problem of missing data by removing data or purposely studying smaller subsets of the data. One of the easiest but not necessarily the most preferable technique is to discard all rows containing missing values. This is referred to as complete-case analysis. Simply removing the missing data introduces two problems: (i) if the units of observation that have missing values differ systematically from the complete cases, the complete-case estimated would be biased, and (ii) if the dataset is large with a high number of predictors the amount of data we discard will likely be a considerable portion of the data [101]. In the complete-case analysis, there is no imputing per se, and it is instead a simple strategy for dealing with missing data.

Single Imputation Instead of removing whole units of observations, the missing data can be filled or imputed with an appropriate replacement value to produce a single complete dataset. This is called single imputation, and there are a variety of such techniques with different complexity. One of the least complex strategies is mean imputation, where the missing values are replaced by the mean of the variable in question. Mean imputation underestimates the variance and distorts the distribution of the imputed variables. Another approach is to fit a regression or classification model on the complete dataset and predict the missing values using the incomplete data as input to the model. The choice of model depends on the missing data, and in recent years deep learning models have also been used as imputation models, in addition to the more traditional linear- and logistic regression models [103]. This imputation strategy

leads to a biased correlation between the predictors, with the imputation strengthening the relationships and reducing variability.

Multiple Imputations To account for the statistical uncertainty in the single imputations, we can create several imputed datasets through multiple imputations (MI) [104]. The technique was developed by Rubin in 1987 and has become the state-of-the-art approach for handling missing data. In MI, we construct m complete datasets, m is usually between 3 and 10, where the missing values are replaced with plausible draws sampled from the predictive distribution of models based on the observed data [103, 105]. The complete datasets are then analyzed using standard complete-data procedures and combined using an MI inference method. The disadvantages with MI compared to single imputation are a higher computational cost and increased workload in analyzing the results [104].

There are various MI methods available. We will focus on the most relevant one to our thesis: multiple imputations with chained equations (MICE). Missing continuous data will be imputed using MICE, with the data being in a wide format.

MICE is based on a modeling strategy where one separately specifies a univariate conditional distribution for every variable that contains missing values, given all the remaining variables, and imputes the variables with missing data iteratively [103]. This strategy is called fully conditional specification and is one of two modeling strategies in MI[1]. The iterative process in MICE can be broken down into four steps [106]:

1. Every missing value is filled using a simple imputation technique, e.g., mean imputation. These imputations can be thought of as "placeholders".

2. The original missing values for one variable, further referred to as "var", are put back into the data, while the other variables with missing data keep the placeholders.

3. Only the units of observations containing observed data for the variable "var" are used in a fitting process. The observed data in "var" is used as the dependent variable and the remaining as independent variables of an appropriate model.

4. The model then imputes the missing data by predicting their values using the fitted model, and the dataset now has one less variable with missing data.

5. Steps 2-4 are repeated for each variable containing missing data. We have completed one cycle or one iteration when the dataset is complete.

6. Steps 2-4 are repeated for a number of cycles, where the imputations are updated at each cycle.

After repeating the steps above a specified number of times, the first imputed dataset out of m datasets is obtained. The process is repeated until we have our desired number of imputed datasets.

The fitted model were originally regression models, with generalized linear models dominating the field, but recent research has shown that specifying the conditional models by classification and regression trees, which we cover in chapter 7, outperforms the traditional models [107]. Hence, our continuous variables will be imputed using MICE with regression trees.

4.2.5 Dimensional Reduction

Dimensional reduction, also called feature reduction, revolves around transforming high dimensional data into a lower dimension, ideally without losing information or variation in the

[1]The second strategy is known as joint modeling. It specifies a joint distribution for the features in the data and generates imputations from the implied conditional distributions of the features with missing data [103].

observed data. Reducing the dimensional space has many potential benefits: it facilitates data visualization and data understanding, reduces the computational cost and time usage during training and requires less storage. On the other hand, information can be lost, and results may be biased if not performed with care. In this thesis, we will reduce the dimension of our data using three different techniques: feature selection, feature aggregation and principal component analysis.

Feature Selection and Aggregation

Feature selection is a procedure where the most important features are selected to create a new subset of the data. Feature aggregation combines multiple features into new ones, which reduces the total number of features in the data. Determining which features are the most important are commonly determined by a variable subset selection algorithm or by applying domain knowledge, creating so-called "ad hoc" features [70]. In this thesis, we will do the latter, and given the large number of such algorithms, we omit the discussion of them entirely, referring the reader to [70] for a more complete coverage of the topic. The aggregation of features must be applied separately to the different data partitions to avoid data leakage.

Principal Component Analysis

Principal component analysis (PCA) assumes that the most relevant information in a signal will reside along the direction with the highest variance. The goal is to identify the direction where the projection of the data points on it has the highest variance. Identifying this direction can be achieved by finding a new basis, constructed as a linear combination of our original data, that emphasizes highly variable directions and reduces redundancy between the basis vectors [108]. Mathematically this translates to finding a transition matrix P that transforms our data into

$$Y = PX^T, \tag{4.5}$$

where X is the $N \times p$ data matrix as before. The rows in P are then a new set of basis vectors for expressing the columns of X^T. The next section describes how we can determine P through the eigenvector decomposition of a symmetric matrix. An equivalent approach is to use the singular value decomposition of X but is omitted for the sake of brevity.

The first step when conducting a PCA is to calculate the sample covariance matrix. From eq. (3.1), the covariance matrix of an arbitrary $m \times n$ matrix B is

$$\begin{aligned} \Sigma(B) &= \mathbb{E}[(B - \mathbb{E}[B])(B - \mathbb{E}[B])^T], \\ &= \mathbb{E}[BB^T] - \mathbb{E}[B]\mathbb{E}[B]^T. \end{aligned} \tag{4.6}$$

Given that the data is from a sample of the population, instead of working with the expression in eq. (4.6), we continue with the sample estimator of the covariance matrix

$$\Sigma(B) = \frac{1}{N-1} BB^T - \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^T, \tag{4.7}$$

where $\tilde{\boldsymbol{\mu}}$ is the sample mean of B. The sample mean is a vector, where each element in $\tilde{\boldsymbol{\mu}}$ is calculated from

$$\tilde{\mu}_j = \frac{1}{N} \sum_{i=1}^{m} B_{ij}, \quad j = 1, \dots n, \tag{4.8}$$

with B_{ij} being the matrix element at the ith row and jth column in B. The data is always assumed to have a sample mean equal to zero when performing a PCA. This leads to the covariance matrix being reduced to

$$\Sigma(B) = \frac{1}{N-1} BB^T. \tag{4.9}$$

This matrix is symmetric, with the off-diagonal elements describing how the different variables co-vary.

If our new basis is to reduce the redundancy between the basis vectors, the covariance matrix in this basis should ideally be a diagonal matrix. The goal then becomes to find a matrix P such that $\Sigma(Y)$ is a diagonal matrix. Inserting the expression for Y from eq. (4.5) into eq. (4.9) we find the relation

$$\Sigma(Y) = \frac{1}{N-1} PAP^T, \tag{4.10}$$

with A being the symmetric matrix $A = X^T X$. From the spectral theorem for symmetric matrices, we know that A is orthogonal diagonalizable by its eigenvectors, meaning it can be expressed as

$$A = VDV^T, \tag{4.11}$$

where V is the matrix formed by the eigenvectors of A and D is a diagonal matrix where the eigenvalues of A constitute the diagonal elements. If we know choose each row in matrix P to be the eigenvectors of A, $P = V^T$, we get the following covariance matrix

$$\Sigma(Y) = \frac{1}{N-1} P(P^T DP)P^T, \tag{4.12}$$

$$= \frac{1}{N-1} D. \tag{4.13}$$

In the last equality we used $P^T = P^{-1}$ for orthogonal matrices . Hence our new basis is the eigenvectors of the covariance matrix, also known as the principal components of X. The trace of D is known as the total variance.

The principal components describe the different directions where the projection of the data points on them has the highest variance, while the eigenvalues give the magnitude. By investigating the magnitude of the eigenvalues, we can identify which direction has the highest variance. This is an important property of the PCA, and from it, we can calculate the explained variance for each principal component. The explained variance is the eigenvalue divided by the total variance.

Reducing the dimensionality of X comes down to how much of the variance we wish to explain with the principal components. In general, we can construct a $N \times k$ matrix, with $k < p$, by selecting the components with the k largest eigenvalues, placing them in the matrix $V_k \in \mathbb{R}^{k \times p}$ and computing $Y_k = V_k X^T$. The choice of k can then be made based on the amount variance we want to explain. For example, if wanting to explain 95% of the variance, we choose k so that

$$\sum_{i=1}^{k} \frac{\lambda_i}{\mathrm{tr}(D)} = 0.95. \tag{4.14}$$

4.3 Model Assessment

Before we delve into the process of selecting an appropriate learning model, we describe how we can assess the model. This is important both in the selection process and when evaluating the final model, given that both procedures involve employing our model on unseen data.

Our model's ability to make correct predictions on new data is evaluated with a performance metric. This metric should not be confused with the loss function, which is vital in the optimization process. However, some performance metrics can be used as a loss function, but this is not universally true for all metrics. In this section, we will introduce three common choices for performance metrics: the mean squared error (MSE), root mean squared error (RMSE) and the mean absolute error (MAE).

Mean Squared Error The MSE of a finite sample of sice N is given by

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(\boldsymbol{x}^{(i)}))^2. \tag{4.15}$$

This MSE is commonly used as a loss function, in addition to being a performance metric. It is differentiable, making it suitable for optimization through gradient descent methods. Because of the squared nature of the metric, it penalizes outliers and small errors, which can lead to an overly negative evaluation of the model. Interpreting its value is not straightforward, given the different scales between input and output.

Root Mean Squared Error To overcome some of the drawbacks of MSE, one can instead use the RMSE, which from the name is naturally given as

$$\text{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(\boldsymbol{x}^{(i)}))^2\right)^{\frac{1}{2}}. \tag{4.16}$$

Here, the input and output scales are the same, and it punishes outliers and minor errors to a smaller degree than the MSE. The RMSE is often preferred when the errors are expected to be normally distributed [109].

Mean Absolute Error Even though the RMSE penalizes outliers to a lesser degree than the MSE, it still punishes variance in the predictions as it weights errors with higher absolute values more than lower errors. The MAE weights all errors equally, and in a finite sample it is calculated from

$$\text{MAE} = \sum_{i=1}^{N}\left|(y_i - \hat{f}(\boldsymbol{x}^{(i)})\right|. \tag{4.17}$$

4.4 Model Selection

Once the data is appropriately processed, we can begin the model selection process. Several algorithms can be applied to the same problem when selecting a model. There is also the possibility of employing the same algorithm with minor variations to the same problem and experiencing different results. How can we choose the best model when there are a number of roads, all of which take us down a slightly different path? To answer this question, we start with a discussion on how bias, variance and model complexity are connected and two possible pitfalls we need to be aware of. The rest of the section is dedicated to how we can circumvent one of these pitfalls.

4.4.1 Bias-Variance Tradeoff

Model complexity, or the number of model parameters in $\hat{f}$, can have a considerable impact on your model's predictive ability. For finite samples, the bias decreases as the model complexity grows, whereas the variance increases. Bias is the error we get from misspecifying the mapping function f from the real underlying function. In the infinite data limit, bias is a measurement of the best possible error on unseen test data, often known as the out-of-sample error. Variance is a measurement of the degree to which the model fluctuates due to finite sample sizes.

Increasing the complexity of the model tailors the model to the specific training data and possibly reduces its generalizability. Here generalizability refers to the model's ability to perform well on unseen data. If the model has been too closely fitted to the data, the model is said to overfit. Overfitting is one of the said pitfalls we need to be wary of. Too low complexity, on the other hand, will also lead to poor generalization, given that the model will not be able to capture

trends in the data. The model is underfitting. In other words, a low model complexity gives a high bias and low variance, which results in underfitting. Increasing the model complexity too far will yield a low bias and high variance, resulting in overfitting. Somewhere in-between these extrema, there is an optimal choice of complexity that yields the lowest possible out-of-sample error.

By investigating the errors computed on the different datasets, we can easily identify if overfitting has occurred. Denoting the error on the training data, also called in-sample error, as E_{in}, and the out-of-sample error as E_{out}. We almost always expect E_{out} to be larger than E_{in} when there is no overfitting. This is not to say that as long as $E_{\text{out}} \geq E_{\text{in}}$, we have not overfit the data. It is simply a rule of thumb that is easily verified, which can help us in our model selection process.

With the increasing computational power and amount of electronic data available, we can construct and train highly complex models compared to the earlier days of machine learning. We are thus more likely to encounter the problem of overfitting instead of underfitting, and a number of techniques have been developed to remedy this issue.

4.4.2 Regularization

Overfitting can always be reduced by increasing the amount of training data. However, acquiring more data is often a time-consuming process, and it is sometimes an impossible way to mitigate the threat of overfitting. By reducing the complexity of a model through regularization, we can circumvent, or at least alleviate, the problem. Some regularization techniques are very algorithm-specific, making it intractable to give a complete overview of these techniques in this thesis. In this section, we restrict ourselves to briefly introducing shrinkage methods, a popular regularization method used in many different supervised learning algorithms. The more algorithm-specific techniques relevant to our specific work are presented in their respective chapters.

Shrinkage Methods

Shrinkage methods imposes a regularizer and a complexity parameter to penalize the complexity of the mapping function $\hat{f}$ in the cost function,

$$\mathcal{C}' = \mathcal{C} + \lambda\Omega(\hat{f}). \tag{4.18}$$

Here $\Omega(\hat{f})$ is the regularizer, which form can vary depending on how we want to regulate the complexity of $\hat{f}$, and $\lambda \geq 0$ is the complexity parameter. Choosing Ω to be either the L_1 or L_2 norm, or a linear combination of the two, of the model parameters are popular choices, with the general L_p norm of a vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ being defined as

$$\|\boldsymbol{x}\|_p = \left(|x_1|^p + \ldots + |x_d|^p\right)^{\frac{1}{p}}. \tag{4.19}$$

This is referred to as L_p-regularization, depending on which value of p is chosen. This form of regularization hinders the model parameters from reaching high values, which is often a characteristic of overfitted models.

4.4.3 Hyperparameters

We briefly described hyperparameters, or nuisance parameters, in sec. 3.2.1 as being parameters that we have no direct inferential interest of. When used in a supervised learning context, hyperparameters are associated with the learning algorithm and are parameters that describe the structure of the model as opposed to the structure of the data. The complexity parameter λ from the previous section is one such parameter. Their values must be assigned before

the training process while the model parameters are obtained during training. Poorly chosen hyperparameters can have a negative impact on the training procedure.

There is no true exact value for these parameters, and therefore they have to be tuned before we can start the training. The process of finding the most desirable hyperparameter values is referred to as hyperparameter optimization. The high number of possible hyperparameter configurations is the root of why there are so many different variations of the same learning algorithm. The theory presented in this section lends inspiration from the work by Bergstra and Bengio on hyperparameter optimization in [110].

As before, we denote a vector of hyperparameters as $\boldsymbol{\lambda}$, where the number of parameters depend on the choice of $\hat{f}$. Instead of decomposing $\boldsymbol{\theta}$ into $(\boldsymbol{\psi}, \boldsymbol{\lambda})$ as we did in subsection 3.2.1, we treat $\boldsymbol{\lambda}$ as an independent set of parameters, with $\boldsymbol{\theta}$ still being the model parameters. Hyperparameter optimization is often performed using cross-validation, which will be covered in sec. 4.4.4. Here the objective is to minimize the mean of errors when the mapping function is applied to several validation sets, for different combinations of hyperparameters, after a number of training procedures. We denote the mapping function of an arbitrary hyperparameter configuration as $\hat{f}_\lambda$. Finding the optimal values for the hyperparameters can thus be expressed as

$$\begin{aligned} \hat{\boldsymbol{\lambda}} &= \arg\min_{\lambda \in \Lambda} \operatorname*{mean}_{X, y \in \mathcal{D}_{\text{val}}} \mathcal{C}(\boldsymbol{y}, \hat{f}_\lambda(X, \boldsymbol{\theta})), \\ &\equiv \arg\min_{\lambda \in \Lambda} \xi(\boldsymbol{\lambda}), \end{aligned} \tag{4.20}$$

where Λ is a set of different combinations of hyperparameter values, referred to as the search space, the function $\xi(\boldsymbol{\lambda})$ is introduced to ease the notation and $\hat{\boldsymbol{\lambda}}$ is the estimated hyperparameter vector. The search space can be made infinitely large, so the strategy is to narrow it down by selecting a finite number of combinations, and evaluate $\xi(\boldsymbol{\lambda})$ on each of them. If we select k different combinations in what is called a trial set, Λ is reduced to $\{\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(k)}\}$, and eq. (4.20) is transformed into

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\lambda \in \{\lambda^{(1)}, \ldots, \lambda^{(k)}\}} \xi(\boldsymbol{\lambda}). \tag{4.21}$$

Ignoring for a moment that hyperparameter optimization is often performed using cross-validation, the general approach for determining the parameters can be summarized in five steps: (i) choose a trial set, (ii) train the model on $\mathcal{D}_{\text{train}}$, (iii) evaluate the model on $\mathcal{D}_{\text{val}}$, (iv) save the results and repeat for a new combination $\boldsymbol{\lambda}^{(j)}$ and (v) compare the results and choose the combination with the best validation score.

Finding the optimal combination can be an exhaustive process for models with multiple hyperparameters, and a manual unsystematic search quickly becomes intractable. However, looking into a limited number of combinations chosen manually can give a quick insight into which regions of the search space are worth exploring. The rest of this section will cover two widely used methods for selecting a trial set.

Grid Search

When performing a grid search for estimating the optimal choice of hyperparameter values, a grid of every possible combination for a chosen number of values is constructed. This grid thus serves as the trial set. Denoting the total number of hyperparameters as H, and the number of values to investigate for each parameter as $L^{(i)}$ for $i = 1, ..., H$, the number of trials to perform in a grid search is $k = \prod_{i=1}^{H} L^{(i)}$. The number of combinations grows exponentially with the number of parameters, and the grid search algorithm is extremely computationally expensive to perform in high-dimensional spaces.

Random Search

Instead of defining a grid from a set of predefined values and performing a systematic grid search, the search space can instead be defined on a bounded domain from which we randomly sample the hyperparameter values. This approach is referred to as a random search, and Bergstra and Bengio [110] showed empirically that a random search is as good, or better, than performing a grid search using a fraction of computation time. Hence, we will apply a random search when tuning hyperparameters in this thesis.

4.4.4 Resampling Techniques

k-Fold Cross Validation

The resampling technique k-fold cross-validation can be applied in the model selection and assessment processes. We include it here since it is a central component in the optimization process for hyperparameters. The method includes the process of training a supervised learning model, a procedure that is explicitly described in the next section, section 4.5.

The core idea behind k-fold cross-validation is that instead of simply evaluating the out-of-sample error on *one* test set, we can evaluate it on k different validation sets. Through this process, we can obtain estimates of the mean and variance of E_{out}. This is achieved by splitting $\mathcal{D}_{\text{train}}$ into k equally sized folds or groups. One fold is chosen as the test set, while the remaining folds serve as the training set. The process is visualized in Figure 4.1.

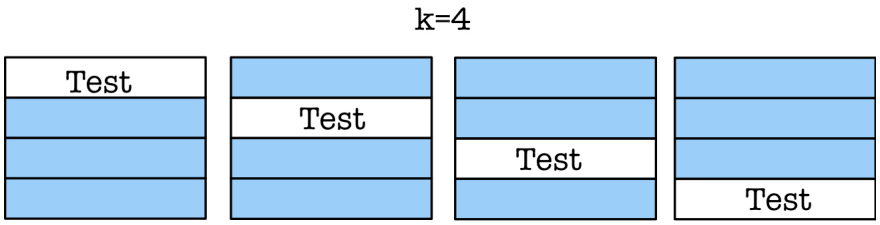

Figure 4.1: Illustration of how a dataset is split into four folds in preparation of a 4-fold cross-validation. The blue folds are used as training data while the one marked "Test" serve as the test set.

Each k fold is used as the test set, meaning that the training procedure is repeated k times. The model trained on the remaining $k-1$ folds is evaluated on the test set, which means we will end up with a $k \times 1$ vector of performance estimates. The mean of this performance vector will give a less biased prediction of the model's performance compared to the case where we use one train/test split for the complete dataset $\mathcal{D}$. When using k-fold cross-validation for hyperparameter optimization, it is important that after the optimal combination of hyperparameters is obtained, the final model is re-trained on the complete training set $\mathcal{D}_{\text{train}}$.

Bootstrapping

Bootstrapping was proposed by Efron in 1979 [111], and is a resampling technique used to obtain accuracy measures for sample estimates. Suppose $\mathcal{D}$ is a dataset containing training data with N points. To perform bootstrap analysis of some sample estimate, we sample B so-called Bootstrap samples $\{\mathcal{D}_1^*, \mathcal{D}_2^*, \cdots, \mathcal{D}_B^*\}$ with N points each, sampled from $\mathcal{D}$ *with* replacement. With a large number of samples we generate an empirical sampling distribution for the statistic of interest. Say we want the sample estimate of some quantity ϕ. With each Bootstrap sample, we compute sample estimates $\{\phi_1^*, \phi_2^*, \cdots, \phi_B^*\}$. From this set it is easy to calculate measures of accuracy for sample estimates, such as confidence intervals.

4.5 Optimization

When we have concluded our model selection process, we can finally begin the training of our selected model by tuning its parameters as described in eq. (4.2). The ultimate goal is to find a minimum or maximum of an objective function, depending on how the supervised learning problem is framed. For some functions, a closed-form solution exists describing all of the extrema, which we can find by setting the gradient equal to zero. Unfortunately, not all functions have a closed-form formulation of their gradient or are not differentiable. The closed-form solution can also be extremely computationally expensive to calculate in other cases. Hence, the need for efficient optimization algorithms arises.

A multitude of optimization algorithms accommodate all of the different scenarios described above. This thesis will restrict itself to iterative algorithms where the gradient is known, specifically, methods based on gradient descent. We will cover four different algorithms: stochastic gradient descent with mini-batches with- and without momentum, ADAM [112] and the subgradient method [113]. Before diving into the theory of the descendants of gradient descent, we introduce the basic theory of their origin.

4.5.1 Gradient Descent

In gradient descent (GD) the model parameters are initialized to some value $\boldsymbol{\theta}_0$, and then iteratively updated in the direction of the steepest descent through the updating scheme

$$\begin{aligned} \boldsymbol{v}_t &= \eta_t \nabla_\theta \mathcal{C}(\boldsymbol{\theta}_t), \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \boldsymbol{v}_t. \end{aligned} \tag{4.22}$$

Here $\nabla_\theta \mathcal{C}(\theta_t)$ is the gradient of the cost function $\mathcal{C}$ with respect to the parameters $\boldsymbol{\theta}$ at iteration step t and η_t is the learning rate, a hyperparameter which controls the step size we move in the direction of the negative gradient.

Determining the learning rate requires several aspects to be taken into consideration: choosing η_t too small is computationally expensive, i. e. the number of iterations needed to converge to the local minimum becomes too large. Depending on the convexity of our cost function, the method will usually converge towards a (possible) local or global minimum. We risk overshooting the minimum with a high learning rate, so choosing η_t wisely is crucial to obtaining satisfactory results. We mention convexity since, for convex functions, any local minimum is a global minimum.

GD has a few drawbacks: computing the gradient for all data points when N is large is highly time-consuming, given that it involves summing over all data points. Furthermore, it lacks stochasticity and treats all directions in the parameter space uniformly, i. e. the learning rate is constant. Since our cost function can have many local minima, GD does not necessarily converge towards the optimal minimum. It is also dependent on the initial conditions $\boldsymbol{\theta}_0$, which, if poorly chosen, can hinder convergence.

4.5.2 Stochastic Gradient Descent with Mini-Batches

To incorporate randomness in GD, the gradient is approximated on subsets of the observed data, called mini-batches, rather than the entire dataset. The batch size is usually significantly smaller than N, typically ranging from a ten to a couple of hundred data points [108].

Denoting the batch size as B, there are B_k mini-batches for $k = 1, ..., N/B$. A full iteration over all the N/B mini-batches, or all the datapoints N, is called an epoch. The number of epochs can be considered a hyperparameter, and it is normal that a model is trained over several epochs. The data points included in each mini-batch are randomly chosen, either with- or without replacement, thus introducing the missing stochasticity, which decreases the likelihood of our fitting algorithm getting stuck in an isolated local minimum. For the case with replacement,

some data points might be picked several times during an epoch, while others may not be picked at all. This approach generally converges more rapidly than the one without replacement [98].

When taking a gradient descent step, the steepest descent direction is now approximated by a mini-batch B_k, as opposed to the full dataset, meaning

$$\nabla_\theta \mathcal{C}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \nabla_\theta \mathcal{C}_i(\boldsymbol{\theta}) \quad \rightarrow \quad \nabla_\theta \mathcal{C}^{MB}(\boldsymbol{\theta}) = \sum_{i \in B_k} \nabla_\theta \mathcal{C}_i(\boldsymbol{\theta}), \tag{4.23}$$

with $\nabla_\theta \mathcal{C}^{MB}(\boldsymbol{\theta})$ now being the mini-batch approximation and $\mathcal{C}(y_i, \hat{f}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})) \equiv \mathcal{C}_i(\boldsymbol{\theta})$. This tackles the time consumption problem that arises with the regular GD method. The updating scheme becomes

$$\begin{aligned} \boldsymbol{v}_t &= \eta_t \nabla_\theta \mathcal{C}^{\mathrm{MB}}(\boldsymbol{\theta}_t), \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \boldsymbol{v}_t. \end{aligned} \tag{4.24}$$

4.5.3 Stochastic Gradient Descent with Mini-Batches and Momentum

A modification which deals with the constant learning rate η_t is to add a momentum term in the updating scheme,

$$\begin{aligned} \boldsymbol{v}_t &= \gamma \boldsymbol{v}_{t-1} + \eta_t \nabla_\theta \mathcal{C}^{\mathrm{MB}}(\boldsymbol{\theta}_t), \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \boldsymbol{v}_t. \end{aligned} \tag{4.25}$$

The learning rate is still constant per se, but the momentum term serves as a memory of the previous time step by adding a fraction $\gamma \in [0, 1)$ of the update vector $\boldsymbol{v}_{t-1}$ from the previous iteration step to the current update vector. Through empirical evidence, default values of both η_t and γ can be found in the literature where they are often set to $(\eta_t, \gamma) = (10^{-3}, 0.9)$ [108].

The momentum term can be interpreted as a weighted mean of recent gradients. The algorithm will move in the direction of small and persistent gradients while suppressing the effects of sudden oscillations. These hypothetical oscillations could arise from the stochasticity introduced in the updating scheme for the stochastic gradient descent (SGD) with mini-batches algorithm. In many cases, this momentum term will increase the convergence rate. However, in some cases, the algorithm can overshoot the local minimum for some choices of γ and η_t, which makes the fitting algorithm oscillate about some local minima and decreases the convergence rate.

4.5.4 ADAM: Adding the Second Moment of the Gradient

In the two preceding updating schemes, the learning rate is constant and, in the case with momentum, an implicit function of the time step t. Ideally, we would want an algorithm that incorporates the curvature of the cost function. When we find ourselves in a flat multi-dimensional landscape, the algorithm takes a larger step than when we are in a steep landscape. This behaviour can be achieved by incorporating the second derivative of the cost function, the Hessian, or its approximation, and normalizing the learning rate by the curvature. However, calculating (or approximating) the Hessian is computationally expensive, so an ideal algorithm implements the adaptive nature of the learning rate without the need for calculating (approximating) the Hessian.

One such algorithm is the ADAM optimizer, which only requires first-order gradients and computes individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients [112]. The nth moment, m_n, of a variable x is defined as $m_n = \mathbb{E}(x^n)$. The algorithm keeps a running average over the gradient and squared gradient through the updating equations

$$\boldsymbol{g}_t = \nabla_\theta \mathcal{C}^{\text{MB}}(\boldsymbol{\theta_t}), \tag{4.26}$$

$$\boldsymbol{m}_t = \beta_1 \boldsymbol{m_{t-1}} + (1-\beta_1)\boldsymbol{g}_t, \tag{4.27}$$

$$\boldsymbol{s}_t = \beta_2 \boldsymbol{s_{t-1}} + (1-\beta_2)\boldsymbol{g}_t^2. \tag{4.28}$$

Here eq. (4.27) and (4.28) are approximations of the first and second moment of the gradient and $\beta_1, \beta_2 \in [0,1)$ are hyperparameters which control the exponential decay rate of these approximations. In their article, Kingma and Lei Ba suggest the following values for the two parameters, $(\beta_1, \beta_2) = (0.9, 0.999)$, and recommend that $\beta_2 > \beta_1$ [112]. Both $\boldsymbol{m}_0$ and $\boldsymbol{s}_0$ are initialized as the zero-vector, meaning that the estimates are biased towards zero. The algorithm thus include to bias-correcting vectors, $\hat{\boldsymbol{m}}_t$ and $\hat{\boldsymbol{s}}_t$,

$$\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-(\beta_1)^t}, \tag{4.29}$$

$$\hat{\boldsymbol{s}}_t = \frac{\boldsymbol{s}_t}{1-(\beta_2)^t}, \tag{4.30}$$

here $(\beta_i)^t$ denotes the hyperparameter β_i to the power t for $i = 1, 2$. The updating scheme for the model parameters is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{s}}_t} + \epsilon}, \tag{4.31}$$

where $\epsilon \sim 10^{-8}$ is a small constant added to prevent division by zero. The adaptive part of the algorithm becomes clearer if we express eq. (4.31) through the variance, $\boldsymbol{\sigma}_t^2 = \hat{\boldsymbol{s}}_t - \hat{\boldsymbol{m}}_t^2$,

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\boldsymbol{\sigma}_t^2 + \hat{\boldsymbol{m}}_t^2} + \epsilon}. \tag{4.32}$$

When the variance is small, i. e. the gradient estimates are consistent through time, the parameters are negatively shifted with a maximum step size of η_t. This limits the maximum step size when we are in steep regions. When the variance is high, $\boldsymbol{\sigma}_t^2 \gg \hat{\boldsymbol{m}}_t^2$, the learning rate is scaled with the standard deviation of the gradient.

4.5.5 Subgradient Method

All of the three optimizations schemes described above assume that the cost function $\mathcal{C}$ is everywhere differentiable. As this is not always the case, an optimization algorithm that does not impose this requirement is needed. One such algorithm is the subgradient method for minimizing non-differentiable convex functions, using subgradients instead of gradients. The subgradient method was developed by Shor [113].

A subgradient of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x}$ is any $\boldsymbol{g} \in \mathbb{R}^n$ such that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{y} - \boldsymbol{x}) \qquad \text{for all } \boldsymbol{y}. \tag{4.33}$$

If f is differentiable, then $\boldsymbol{g}$ is uniquely equal to the gradient of f. Non-differentiable functions can have a infinite number of subgradients at a non-differentiable point, and from the definition in eq. (4.33) these are vectors that underestimate the function f. The set of all subgradients of f at $\boldsymbol{x}$ is called the subdifferential of f at $\boldsymbol{x}$, and is denoted $\partial f(\boldsymbol{x})$. The update scheme for the subgradient method is similar to the one of GD, being given as

$$\begin{aligned} \boldsymbol{v}_t &= \eta_t \boldsymbol{g}_t, \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \boldsymbol{v}_t. \end{aligned} \tag{4.34}$$

Here g_t is any subgradient of f at iteration t, i. e. $g_t \in \partial f(\boldsymbol{\theta}_t)$, with $\eta_t > 0$ still being the learning rate. As in GD, the parameters are initialized to some value $\boldsymbol{\theta}_0$. As opposed to the previous optimization schemes, the subgradient method is not a decent method; it is therefore common to save the smallest function value so that in each iteration we set

$$f_t^{(\text{best})} = \min\{f(\boldsymbol{\theta}_1), \ldots, f(\boldsymbol{\theta}_t)\}. \tag{4.35}$$

When applying the subgradient method the learning rate is set before the algorithm starts running, meaning that there is no tuning of η. There exists several rules to decide the learning rate, but using either a constant, $\eta_t = \eta$, or a constant reduction term, $\eta_t = \gamma / \|g_t\|_2$ for some $\gamma > 0$, the subgradient method is guaranteed to converge to some optimal value such that

$$\lim_{t \to \infty} f_t^{(\text{best})} - f^* < \epsilon, \tag{4.36}$$

where ϵ is some threshold and f^* is the optimal value for the given problem [114].

4.5.6 Summary

We summarize this section by including a table of the different GD algorithms and their hyperparameters with corresponding default values. The summary can be found in Table 4.3.

Table 4.3: The table displays the different hyperparameters and their default values for three gradient descent methods: stochastic gradient descent (SGD) with- and without momentum and the ADAM algorithm.

Algorithm	Hyperparameters (Default Values)		
SGD	η (10^{-3})		
SGD with Momentum	η (10^{-3})	γ (0.9)	
ADAM	η (10^{-3})	β_1 (0.9)	β_2 (0.999)
Subgradient Method	η (10^{-3})		

After the optimization procedure is complete, the final step is to evaluate our trained model on $\mathcal{D}_{\text{test}}$ using the chosen performance metric.

Chapter 5

Linear Models

Linear regression models are frequently used in many fields of science, and among them are both physics education research and psychology. Regression is a set of statistical methods for estimating the relationship between a dependent variable and a set of independent variables that are related in a non-deterministic fashion [115]. The relationship between the variables is expressed through a mathematical model, and a suitable estimation method is chosen to estimate the model parameters, also called regression coefficients. Regression deals with continuous data and is a supervised learning technique.

This thesis will use multiple linear regression models to represent conventional statistical methods in the field of psychology. We will also fit a regularized linear model, an elastic net, to the data. This chapter is therefore dedicated to developing the theory behind these types of models and their estimation methods. We start the chapter by introducing the multiple linear regression model in section 5.1, followed by regularized linear models.

5.1 Multiple Linear Regression

Suppose we have a dataset $\mathcal{D}$ on same form as the one in sec. 4.1, $\mathcal{D}(X, \boldsymbol{y}) = \{(\boldsymbol{x}^{(i)}, y_i)\}_{i=1}^{N}$. In multiple linear regression we assume that the data is generated from $y_i = f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}_{\text{true}}) + \varepsilon_i = \boldsymbol{\beta}_{\text{true}}^T \boldsymbol{x}^{(i)} + \varepsilon_i$, with the true parameters denoted as $\boldsymbol{\beta}_{\text{true}}$. Hence, a multiple linear regression model predicts a deterministic output $\hat{y}_i$ via the model

$$\hat{y}_i = \hat{f}(\boldsymbol{x}^{(i)}; \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j, \tag{5.1}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T \in \mathbb{R}^{p+1}$ being the regression coefficients. The term β_0 is known as the intercept, or bias, and is the mean value of $\hat{y}_i$ when all of the predictors are equal to zero. For estimating the intercept, an all-ones column vector can be added as the first column in X, making X a matrix of dimensions $N \times (p+1)$, often called the design matrix. In matrix notation, the model in eq. (5.1) can be expressed as

$$\hat{\boldsymbol{y}} = X\boldsymbol{\beta}, \tag{5.2}$$

with $\hat{\boldsymbol{y}} = (\hat{y}_1, ..., \hat{y}_N)^T \in \mathbb{R}^N$.

The word multiple in "multiple linear regression models" refers to the fact that there are several independent variables. In the case where there is only one independent variable, the model is reduced to a simple linear regression model.

5.1.1 Interpreting Regression Coefficients

One of the reasons why linear regression models are a popular choice across scientific fields is their interpretability. The regression coefficients' values tell us the size of the effect each

independent variable has on the dependent variable when the independent variable experiences a one-unit increase while the remaining variables are held constant. The sign indicates whether the effect increases or decreases in the dependent variable.

The standard error of a regression coefficient can determine if the coefficient is estimated precisely, and the smaller the error, the more precise the estimate. The standard error can again be used to determine the significance of the coefficient by calculating its p-value, which can be considered one of the main goals when running a regression analysis. However, a coefficient can be statistically significant and still have a negligible effect size on the dependent variable.

5.1.2 Assumptions

Using eq. (5.2) to describe a relationship between data requires us to make some assumptions about the predictors, the response and their errors. This paragraph will give an overview of the assumptions made by the estimation method described in the next section, relying on the theory presented in [101, p. 45].

1. *Linearity*: The relationship between the model parameters and the independent variables is modeled linearly with respect to the parameters, which translates to that the predicted output can be expressed as a linear combination of the predictors and regression coefficients.

2. *Independence of Errors*: The multivariate regression model assumes that the errors in the response are independent of each other, i. e., uncorrelated.

3. *Equal Variance of Errors*: Also called homoscedasticity, is the assumption that the variance of the errors does not depend on the values of the predictors. The absence of homoscedasticity is called heteroscedasticity.

4. *Linear Independence in Predictors*: In the case where a predictor is a linear combination of some other predictors, the data is said to be perfect multicollinear. Some methods accept a degree of multicollinearity, but if there is a linear relationship between two or more predictors, no unique solution can be obtained. This is related to the rank of X, where rank is the number of linearly independent columns in X. If $\text{rank}(X) = p$, the solution is unique since all the columns are independent. In the case where $p > N$, no unique solution exists.

5.1.3 Estimation with Ordinary Least Squares

One of the most common estimation methods for multiple regression models is ordinary least squares (OLS), which seeks to minimize the cost function

$$\mathcal{C}(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2. \tag{5.3}$$

The cost function is the squared L_2-norm of the residuals, also called the residual sum of squares. Formally, this means we seek a solution vector $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ such that

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2. \tag{5.4}$$

Differentiation of the squared L_2 norm with respect to the parameters yields the solution vector

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(X^T X\right)^{-1} X^T \boldsymbol{y}. \tag{5.5}$$

We have assumed $X^T X$ to be invertible, which is usually the case when the number of datapoints N exceeds the number of features p such that $N \gg p$ [108].

5.1.4 Estimation with Generalized Least Squares

When the assumption about the independence of errors is violated, OLS is no longer the suitable choice for estimating the regression coefficients in eq. (5.2). Differing variance in errors is the same as having a covariance matrix that is no longer diagonal. Meaning that the errors are now distributed as $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, V)$, with V being the non-diagonal covariance matrix. Estimating this model is done with the generalized least squares (GLS), assuming V is known.

All covariance matrices are symmetric and positive semi-definite, meaning that it can be factored into $V = QQ^T$ using the Cholesky decomposition, for some invertible matrix Q. Mulitplying the true model with Q^{-1} yields the transformed equation

$$\boldsymbol{y}' = X'\boldsymbol{\beta} + \boldsymbol{\varepsilon}', \tag{5.6}$$

with $\boldsymbol{y}' \equiv Q^{-1}\boldsymbol{y}$, $X' \equiv Q^{-1}X$ and $\boldsymbol{\varepsilon}' \equiv Q^{-1}\boldsymbol{\varepsilon}$. If $\boldsymbol{\varepsilon}' \sim \mathcal{N}(0, V')$, where V' is a diagonal matrix, the assumption about uncorrelated errors is met, and we can employ OLS to acquire $\hat{\boldsymbol{\beta}}_{\text{GLS}}$. The expectation value and covariance matrix of $\boldsymbol{\varepsilon}'$ are

$$\mathbb{E}(Q^{-1}\boldsymbol{\varepsilon}) = Q^{-1}\mathbb{E}(\boldsymbol{\varepsilon}) = 0, \tag{5.7}$$

$$\begin{aligned}\text{Cov}(Q^{-1}\boldsymbol{\varepsilon}) &= Q^{-1}\text{Cov}(\boldsymbol{\varepsilon})(Q^{-1})^T, \\ &= Q^{-1}V(Q^{-1})^T = \sigma^2 I,\end{aligned} \tag{5.8}$$

where we in the last equality used the decomposition of V. The resulting matrix is a diagonal matrix, with I denoting the identity matrix and σ^2 here being a variance vector. The OLS solution of eq. (5.6) is thus

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left((X')^T X'\right)^{-1}(X')^T\boldsymbol{y}'. \tag{5.9}$$

Inserting the primed variables into the equation above yields the GLS solution vector,

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (X^T V^{-1} X)^{-1} X^T V^{-1}\boldsymbol{y}. \tag{5.10}$$

5.1.5 Reformulating as a Maximum Likelihood Estimation Problem

The solution vector for the two linear regression problems above can also be obtained through maximum likelihood estimation. This section is dedicated to deriving the solution vector for OLS using the MLE formalism from section 3.2, lending inspiration from the treatment the same topic is given in [108].

Our regression model in eq. (5.2) can be defined by a conditional probability that is both dependent on the data and some parameters $\boldsymbol{\theta}$, which can be expressed as

$$p(y_i|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}) = \mathcal{N}\left(y_i|\mu(\boldsymbol{x}^{(i)}), \sigma^2(\boldsymbol{x}^{(i)})\right), \quad \forall i = 1, ...N. \tag{5.11}$$

Here $\mu(\boldsymbol{x}^{(i)}) = \boldsymbol{\beta}^T\boldsymbol{x}^{(i)}$ is the mean, which in our case is a linear function of $\boldsymbol{x}$. The variance σ^2 is fixed, i. e. $\text{Var}(\boldsymbol{x}^{(i)}) = \sigma^2$, hence the parameters $\boldsymbol{\theta}$ are $(\boldsymbol{\beta}, \sigma^2)$. The log-likelihood for the model in eq. (5.1) is

$$\begin{aligned}l(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \boldsymbol{\beta}^T\boldsymbol{x}^{(i)}\right)^2 - \frac{N}{2}\log\left(2\pi\sigma^2\right), \\ &= -\frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + C,\end{aligned} \tag{5.12}$$

with $C \equiv -N/2\log\left(2\pi\sigma^2\right)$ being a constant. From eq. (5.12) it becomes evident that maximizing the log-likelihood will yield the same result as minimizing eq. (5.4).

5.2 Regularized Linear Models

To prevent overfitting in linear models, it is common to impose a regularization term to the cost function in eq. (5.3) that penalizes large weights. By reducing the size of the weights, we are reducing the parameter space, which again decreases our chance of overly adapting our linear model to the data.

5.2.1 Ridge Regression

In the case where the number of variables exceeds the number of observations, $p > N$, a unique solution may not exists due to singularities in the matrix $X^T X$. When this is the case, small changes in X can lead to large changes in $X^T X$, so by introducing small perturbations to $X^T X$ we can overcome this potential pitfall. By imposing a L_2-regularizer to the cost function in eq. (5.3) we circumvent the problem with singularity and also prevent overfitting by constraining the model parameters. This is known as Ridge regression and was proposed by Hoerl and Kennard in 1970 [116].

The cost function in ridge regression is

$$\mathcal{C}(\lambda, \boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{5.13}$$

with $\lambda > 0$ being a complexity parameter. The solution vector is found from the minimization problem

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2. \tag{5.14}$$

An equivalent way of expressing the minimization problem is

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}}(t) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}:\|\beta\|_2^2\leq t}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2. \tag{5.15}$$

Here the constraint on the parameters are made explicitly. There is a one-to-one correspondence between λ and t, meaning that for any $t \geq 0$ and solution vector $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$ in eq. (5.15), there exists a $\lambda \geq 0$ such that $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$ solves eq. (5.14) [117].

As with OLS, an analytical expression of $\hat{\boldsymbol{\beta}}_{\text{Ridge}}$ is obtainable by differentiating eq. (5.13) with respect to the model parameters:

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = \left(X^T X + \lambda I_{p\times p}\right)^{-1} X^T \boldsymbol{y}. \tag{5.16}$$

Here $I_{p\times p}$ denotes the $p \times p$ identity matrix.

5.2.2 Lasso Regression

The Lasso shrinkage method also penalizes the model parameters, but uses the L_1-norm. The distinction may seem subtle, but has some major implications for the final solution. The Lasso estimate is given by

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{5.17}$$

Like with ridge regression, eq. (5.17) can be expressed as a constrained minimization problem,

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}}(t) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}:\|\beta\|_1\leq t}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2. \tag{5.18}$$

Lasso regression performs a kind of continuous variable selection, given that small values of t lead some coefficients to be exactly zero [117]. The L_1-norm is not everywhere differentiable, hence there is no closed-form expression of $\hat{\boldsymbol{\beta}}_{\text{Lasso}}$.

Obtaining a solution of eq. (5.17) can be achived by utilizing an optimization scheme for non-differentiable cost function, such as the subgradient method described in section 4.5. Recalling that the subgradient method was only applicable to convex functions, we must prove that $\mathcal{C}_{\text{Lasso}}$ is convex. There are several definition of convexity, and to prove that the cost function in Lasso regression is convex we will use two separate definitions. The first one states that a function f is said to be convex if its Hessian is a positive semi-definite matrix. A symmetric matrix A is said to be positive semi-definite if the relation

$$\boldsymbol{u}^T A \boldsymbol{u} \geq 0, \tag{5.19}$$

is fulfilled for all vectors $\boldsymbol{u} \neq \mathbf{0}$. The second definition states that a real function f is convex on the interval I if and only if

$$f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) \leq tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2), \tag{5.20}$$

for all $t \in [0,1]$ and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in I$.

Decomposing the cost function into

$$\mathcal{C}_{\text{Lasso}} = \mathcal{C}_{\text{OLS}} + \lambda \|\boldsymbol{\beta}\|_1, \tag{5.21}$$

and calculating the Hessian for $\mathcal{C}_{\text{OLS}}$ as

$$\nabla^2 \mathcal{C}_{\text{OLS}} = X^T X. \tag{5.22}$$

The matrix $X^T X$ is symmetric, so inserting the Hessian into the inequality in eq. (5.19) yields

$$\boldsymbol{u}^T \left(\nabla^2 \mathcal{C}_{\text{OLS}}\right) \boldsymbol{u} = \boldsymbol{u}^T X^T X \boldsymbol{u} = \|X\boldsymbol{u}\|^2 \geq 0, \tag{5.23}$$

which shows that $\mathcal{C}_{\text{OLS}}$ is a convex function. Now using the second definition of convexity on the second term in eq. (5.21) gives us the inequality

$$\begin{aligned} |t\boldsymbol{x_1} + (1-t)\boldsymbol{x_2}| &\leq |t\boldsymbol{x_1}| + |(1-t)\boldsymbol{x_2}|, && \triangleright \text{ Triangle Inequality} \\ &= |t||\boldsymbol{x_1}| + |(1-t)||\boldsymbol{x_2}|, \\ &= t|\boldsymbol{x_1}| + (1-t)|\boldsymbol{x_2}|. \end{aligned} \tag{5.24}$$

The absolute value of a vector is thus a convex function. Given that the cost function in Lasso regression is made up of two convex functions, and the sum of two convex functions simply yields a new convex function, the subgradient method can be applied to obtain an estimate for $\hat{\boldsymbol{\beta}}_{\text{Lasso}}$.

The subgradient for the cost function in Lasso is

$$\overset{\text{sub}}{\nabla} \mathcal{C}_{\text{Lasso}} = \nabla \mathcal{C}_{\text{OLS}} + \lambda \overset{\text{sub}}{\nabla} \|\boldsymbol{\beta}\|_1. \tag{5.25}$$

Remember that a subgradient of a function at the point where it is non-differentiable can be any vector that underestimates the function in that point. The L_1-norm is non-differentiable in the point $\boldsymbol{\beta} = \mathbf{0}$, so any vector in the interval $\boldsymbol{\beta} \in [-1,1]^p$ is a subgradient. This can easily be seen by applying the definition of a subgradient from eq. (4.33) to univariate case $|\beta|$. The subgradient of the L_1-norm is thus

$$\overset{\text{sub}}{\nabla} \|\boldsymbol{\beta}\|_1 = \begin{cases} 1 & \boldsymbol{\beta} > \mathbf{0} \\ \boldsymbol{\beta} \in [-1,1]^p & \boldsymbol{\beta} = \mathbf{0} \\ -1 & \boldsymbol{\beta} < \mathbf{0} \end{cases} \tag{5.26}$$

We can conveniently choose to look at the vector in $\boldsymbol{\beta} = \mathbf{0}$, given that $\mathbf{0} \in [-1,1]^p$, which reduces the subgradient in eq. (5.26) to the sign operator. The Lasso estimate is then found from applying the subgradient method with the subgradient

$$\overset{\text{sub}}{\nabla} \mathcal{C}_{\text{Lasso}} = \nabla \mathcal{C}_{\text{OLS}} + \lambda \cdot \text{sgn}(\boldsymbol{\beta}). \tag{5.27}$$

5.2.3 Elastic Net

The elastic net was proposed in 2005 by Zou and Hastie [118], and is a linear combination of both the Ridge and Lasso penalty,

$$\mathcal{C}(\lambda_1, \lambda_2, \boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1. \tag{5.28}$$

The elastic net estimate, $\hat{\boldsymbol{\beta}}_{\text{EN}}$, is then

$$\hat{\boldsymbol{\beta}}_{\text{EN}}(\lambda_1, \lambda_2) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1, \tag{5.29}$$

which can also be expressed as the constrained problem

$$\hat{\boldsymbol{\beta}}_{\text{EN}}(t) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}:(1-\alpha)\|\beta\|_1+\alpha\|\beta\|_2^2\leq t}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2, \tag{5.30}$$

where $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$. When $\alpha = 1$, eq. (5.30) is reduced to Ridge regression, and $\alpha = 0$ gives the Lasso penalty.

The cost function of the elastic net in eq. (5.28) can be decomposed into two terms, with the first being $\mathcal{C}_{\text{Ridge}}$ and the second being the Lasso penalty. The cost function in Ridge regression is a convex function, with its Hessian being $\nabla^2\mathcal{C}_{\text{Ridge}} = X^TX + \lambda_2 I$ which always satisfy the inequality from eq. (5.19):

$$\boldsymbol{u}^T\left(\nabla^2\mathcal{C}_{\text{Ridge}}\right)\boldsymbol{u} = \boldsymbol{u}^T(X^TX + \lambda I)\boldsymbol{u} = \|Xu\|^2 + \lambda\|\boldsymbol{u}\|^2 \geq \lambda\|\boldsymbol{u}\|^2 > 0. \tag{5.31}$$

We note that the cost function in Ridge is strictly convex given that eq. (5.31) is always greater than zero. With the L_1-norm being a convex function, the solution vector $\hat{\boldsymbol{\beta}}_{\text{EN}}$ can be obtained by the subgradient method with the subgradient

$$\overset{\text{sub}}{\nabla}\mathcal{C}_{\text{EN}} = \nabla\mathcal{C}_{\text{Ridge}} + \lambda \cdot \text{sgn}(\boldsymbol{\beta}). \tag{5.32}$$

The elastic net also performs a continuous variable selection, as with Lasso. This makes both methods preferable over Ridge, given the desire of scientists to produce parsimonious models. Ridge regression always keeps all the predictors in the model. The elastic net is shown to have a "grouping-effect" that is not evident in the Lasso regression [118]. This means that if a group of variables is highly pairwise correlated, the Lasso tends to only choose one of the variables, i. e. the Lasso assigns very different coefficients β_i to similar values. In the extreme case where two variables are identical, the regression method should assign identical coefficients to the values. However, this behaviour is only ensured if $\mathcal{C}$ is strictly convex [118]. The Ridge penalty in the elastic net satisfies this criterion, explaining why a grouping effect is observed in the elastic net and not in Lasso regression.

Chapter 6

Neural Nets

Neural nets (or neural networks) are non-linear machine learning models for supervised learning. They have gained massive popularity in recent years and are recognized as one of the most powerful and widely used learning models [108]. Neural nets are versatile models, meaning that they can be constructed into a high number of different architectures. For that reason, it is helpful to divide the different types of networks into categories: general-purpose neural networks for supervised learning, networks designed for tasks involving image processing, architectures designed for sequential data such as text or speech recognition and neural networks for unsupervised learning. This thesis will apply networks that fit the first category description, specifically, a multilayer perceptron.

This chapter starts by introducing the fundamental structure of a network, followed by the mathematical notation needed to understand the underlying workings of such a network. In section 6.3 we describe how the network is trained through the backpropagation algorithm. In the following section, we list the complete algorithm along with our choice of cost function and top layer activation function. In section 6.4 we discuss activation functions for the hidden layers, followed by a description on how to initialize the model parameters in section 6.5. Finally, in section 6.6 we discuss some popular regularization techniques for neural nets.

6.1 Structure

The basic building blocks of a multilayer perceptron are a set of layers, each of which has a set of neurons (or nodes). These layers are usually divided into three categories:

1. The input layer, here a real observable $\boldsymbol{x} \in \mathbb{R}^p$ is fed to the network.

2. Hidden layers, where the processing of the information happens.

3. The output or top layer, where a response $\hat{y}$ is calculated.

The fundamental building blocks in a neural net are called neurons (or nodes), which take a vector $\boldsymbol{x} \in \mathbb{R}^p$ as input and output a scalar through a non-linear activation function. A neural net consists of many such neurons stacked together into layers. Between any two adjacent layers, a set of weights connects a neuron from one layer to every neuron in the adjacent layer. Typically, a bias is added to each neuron in every layer to help facilitate activation of the neuron. The weights and biases are considered the model parameters for a neural net. Figure 6.1 depicts the network architecture of a multilayer perceptron with two hidden layers and one single output.

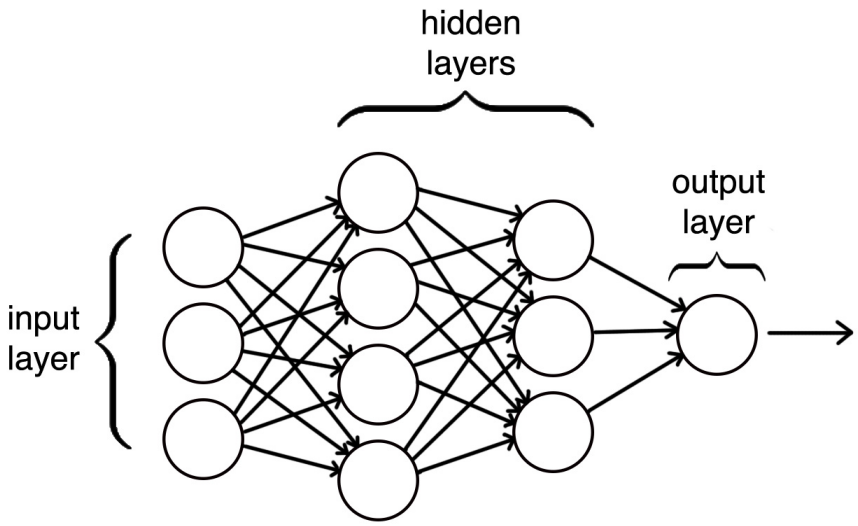

Figure 6.1: The figure shows a multilayer perceptron with one input layer, two hidden layers and one output layer with a single output. Here the hidden layers have an unequal number of nodes. The figure is adapted from Mehta et. al [108].

The power of neural networks can somewhat be explained through the universal approximation theorem, which states that neural networks with a single hidden layer can approximate any continuous function to any desired precision [119]. A higher number of hidden layers are also associated with being able to learn more complex features from the data [108]. Since we are free to add as many hidden layers as we want, neural networks can thus approximate more complicated functions compared to linear models.

6.2 Basic Mathematical Formalism

Let $\mathcal{D}$ denote the data as in the previous chapters. Instead of assuming one output per input vector $\boldsymbol{x}^{(i)}$, we are in this chapter going to be working with multiple outputs, simply because the theory is somewhat identical for the two cases. Here we assume m number of outputs. Further, denote a layer in the neural network by l. Let there be L layers in total such that $l \in \{1, 2, ..., L\}$. Denote the activation of neuron j in layer l by a_j^l and its corresponding bias as b_j^l. Finally, let W_{jk}^l be the weights connecting neuron k in layer $l-1$ to neuron j in layer l. The activation a_j^l is modelled through the (in general) non-linear equation

$$a_j^l = \sigma\left(\sum_k W_{jk}^l a_k^{l-1} + b_j^l\right) \equiv \sigma(z_j^l), \tag{6.1}$$

with z_j^l being the weighted sum

$$z_j^l = \sum_k W_{jk}^l a_k^{l-1} + b_j^l, \tag{6.2}$$

and $\sigma(\cdot)$ is some (possibly) non-linear function colloquially called the activation function, which in this context should not be confused with the statistical measure of dispersion known as the standard deviation. In sec. 6.4 we discuss common choices for σ. For the input layer, i. e. $l = 1$, eq. (6.1) specializes to

$$a_j^1 = \sigma\left(\sum_{k=1}^{p} W_{jk}^1 x_k + b_j^1\right), \tag{6.3}$$

where x_k for $k = 1, ..., p$ are elements in the input vector $\boldsymbol{x}$.

The number of outputs is determined by the number of neurons in the top layer. In the case with one output, $j = 1$ in eq. (6.1). With multiple outputs, j would simply run from one up to m.

6.2.1 Brief Note on Model Complexity

The model complexity of a neural net boils down to how many parameters the model has. Let P denote the number of model parameters and let n be the number of nodes in each layer, assuming that all the hidden layers have an equal number of nodes. Furthermore, p is the number of features in the input vector. The total number of parameters can then be expressed as:

$$P = n(p+1) + (L-2)n(n+1) + m(n+1), \tag{6.4}$$

where the first term corresponds with the input layer, the second term describes all hidden layers which are simply connected to another hidden layer, and the last term is the parameters needed in the top layer. The plus one in all terms comes from the bias connected to each layer. If all layers have the same number of nodes, we note that increasing the number of hidden layers increases the model complexity far more than if we increase the number of nodes in each layer.

6.3 The Backpropagation Algorithm

Based on eq. (6.1), a neural network can be represented by a composite function, a function of the form

$$f(\boldsymbol{x}) = (f_L \circ f_{L-1} \circ \cdots \circ f_1)(\boldsymbol{x}), \tag{6.5}$$

whose gradient carries a computational complexity that makes a direct calculation intractable, or at the very least very computationally expensive. The nested structure of the network makes the cost function indirectly dependent on all neurons in the net. A clever circumvention is to create a set of equations defining a recursive relationship such that knowledge of the error in any layer $l+1$ permits us to deduce the error in layer l. This is known as the backpropagation algorithm [120]. The algorithm is split into two parts, a forward pass and a backward pass. The forward pass calculates all neurons' activation in all layers, while the backward pass propagates the error recursively through the network.

The core of the backward pass is based on the chain rule for partial differentiation and can be summarized with four equations. Our final goal is to obtain the partial derivative of the cost function with respect to the model's parameters, the weights and biases, i. e. $\partial\mathcal{C}/\partial W^l_{jk}$ and $\partial\mathcal{C}/\partial b^l_j$. When they are obtained, we can perform one optimization step, and in this thesis, we will be utilizing the Adam optimizer described in section 4.5.

We denote the error at neuron j in layer l as Δ^l_j. The first of the four equations is the error at layer L with respect to the weighted sum z^L_j, which is defined as

$$\Delta^L_j = \frac{\partial\mathcal{C}}{\partial z^L_j} = \frac{\partial\mathcal{C}}{\partial a^L_j}\sigma'(z^L_j). \tag{I}$$

Here $\sigma'(\cdot)$ is the derivative of the activation function with respect to its input. For the remaining $L-1$ layers, the error with can be computed in a similar manner,

$$\Delta^l_j = \frac{\partial\mathcal{C}}{\partial z^l_j} = \frac{\partial\mathcal{C}}{\partial a^l_j}\sigma'(z^l_j). \tag{6.6}$$

We notice that if we compute the partial derivative of the cost function with respect to the bias b^l_j we end up with the same expression as for Δ^l_j,

$$\frac{\partial\mathcal{C}}{\partial b^l_j} = \frac{\partial\mathcal{C}}{\partial z^l_j}\frac{\partial b^l_j}{\partial z^l_j} = \Delta^l_j, \tag{II}$$

given that $\partial b_j^l/\partial z_j^l = 1$. This is the second of the four equations needed to describe the backpropagation algorithm.

We obtain the third equation by applying the chain rule to Δ_j^k. Given the composite nature of a neural net, the cost function is automatically a composite function as well. Hence, the activations for layer l does not appear explicitly in the cost function before we evaluate the subsequent layer $l+1$. With that in mind, the partial derivative of $\mathcal{C}$ with respect to z_j^l can be written as

$$\begin{aligned}\Delta_j^l = \frac{\partial \mathcal{C}}{\partial z_j^l} &= \sum_k \frac{\partial \mathcal{C}}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l}, \\ &= \sum_k \Delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_j^l}. \end{aligned} \tag{6.7}$$

The partial derivative in eq. (6.7) can simply be expressed as

$$\begin{aligned}\frac{\partial z_k^{l+1}}{\partial z_j^l} &= \frac{\partial z_k^{l+1}}{\partial a_j^l} \sigma'(z_j^l), \\ &= W_{kj}^{l+1} \sigma'(z_j^l),\end{aligned}$$

where we in the last line used

$$\frac{\partial z_k^{l+1}}{\partial a_j^l} = \frac{\partial}{\partial a_j^k}\left(\sum_s W_{ks}^{l+1} a_s^l + b_k^{l+1}\right) = \sum_s W_{ks}^{l+1} \Delta_{sj} = W_{kj}^{l+1},$$

with δ_{sj} being the Kronecker delta. Inserting this result into eq. (6.7) yields the third equation

$$\Delta_j^l = \left(\sum_k \Delta_k^{l+1} W_{kj}^{l+1}\right) \sigma'(z_j^l). \tag{III}$$

The final equation comes from taking the partial derivative of $\mathcal{C}$ with respect to the weight W_{jk},

$$\frac{\partial \mathcal{C}}{\partial W_{jk}^l} = \frac{\partial \mathcal{C}}{\partial z_j^l} \frac{\partial z_j^l}{\partial W_{jk}^l} = \Delta_j^l a_k^{l-1}. \tag{IV}$$

We have now derived all of the four equations needed to perform the backpropagation algorithm, for a single data point. When listing the complete algorithm it is convenient to do so using matrix notation. Rewriting the equations from index to matrix notation requires us to introduce the Hadamard product, which is elementwise multiplication of two vectors of the same dimension. The Hadamard product of two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is denoted $\boldsymbol{u} \odot \boldsymbol{v}$, where the ith component is simply $(\boldsymbol{u} \odot \boldsymbol{v})_i = u_i v_i$. The four equations can now be rewritten into

$$\Delta_j^L = \frac{\partial \mathcal{C}}{\partial a_j^L} \sigma'(z_j^L) \quad \rightarrow \quad \boldsymbol{\Delta}^L = \frac{\partial \mathcal{C}}{\partial \boldsymbol{a}^L} \odot \sigma'(\boldsymbol{z}^L), \tag{I}$$

$$\frac{\partial \mathcal{C}}{\partial b_j^l} = \Delta_j^l \quad \rightarrow \quad \frac{\partial \mathcal{C}}{\partial \boldsymbol{b}^l} = \boldsymbol{\Delta}^l, \tag{II}$$

$$\Delta_j^l = \left(\sum_k \Delta_k^{l+1} W_{kj}^{l+1}\right) \quad \rightarrow \quad \boldsymbol{\Delta}^l = \left((W^{l+1})^T \boldsymbol{\Delta}^{l+1}\right) \odot \sigma'(\boldsymbol{z}^l), \tag{III}$$

$$\frac{\partial \mathcal{C}}{\partial W_{jk}^l} = \Delta_j^l a_k^{l-1} \quad \rightarrow \quad \frac{\partial \mathcal{C}}{\partial W^l} = \boldsymbol{\Delta}^l \otimes \boldsymbol{a}^{l-1}. \tag{IV}$$

The symbol $\otimes$ denotes the outer product between two vectors in the final equation.

Before summarizing the four equations in a complete algorithm, we need an explicit expression for the error at the top layer.

6.3.1 Top Layer Activation and Choice of Cost Function

The rate at which a neural network learns is highly dependent on the combination of top layer activation and cost-function. For regression, a natural choice for top layer activation, and the one that we are going to be working with in this thesis, is simply the linear relationship

$$a_j^L = z_j = \sum_k W_{jk}^L a_k^{L-1} + b_j^L. \tag{6.8}$$

In this case, choosing the cost-function (for a single data point) to be

$$\mathcal{C} = \frac{1}{2}\sum_j (y_j - a_j^L)^2, \tag{6.9}$$

gives the following error at layer L

$$\Delta_j^L = \frac{\partial \mathcal{C}}{\partial z_j^L} = \frac{\partial \mathcal{C}}{\partial a_j^L}\frac{\partial a_j^L}{\partial z_j^L} = a_j^L - y_j, \tag{6.10}$$

which is the same as

$$\boldsymbol{\Delta}^L = \boldsymbol{a}^L - \boldsymbol{y}, \tag{6.11}$$

in matrix notation. The factor 2^{-1} in eq. (6.9) is added for convenience when computing the derivative. Without the factor, the cost function chosen is simply the residual square of sums.

We are now ready to list the backpropagation algorithm in algorithm 1.

Algorithm 1 Backpropagation algorithm in matrix notation. The algorithm consists of two distinct parts: a feed-forward pass and a backward pass.

procedure FEED-FORWARD($\boldsymbol{x}$)
  $\boldsymbol{a}^0 = \boldsymbol{x}$ ▷ Initialize input
  for $l = 1, ..., L$ **do**
    $\boldsymbol{a}^l \leftarrow \sigma(W^l \boldsymbol{a}^{l-1} + \boldsymbol{b}^l)$
procedure BACKWARD PASS($\boldsymbol{y}$)
  $\boldsymbol{\Delta}^L = \boldsymbol{a}^L - \mathbf{y}$ ▷ eq. (I)
  for $l = L-1, ..., 1$ **do**
    $\boldsymbol{\Delta}^l \leftarrow \left((W^{l+1})^T \boldsymbol{\Delta}^{l+1}\right) \odot \sigma'(\boldsymbol{z}^l)$ ▷ eq. (III)
    $\partial \mathcal{C}/\partial \boldsymbol{b}^l \leftarrow \boldsymbol{\Delta}^l$ ▷ eq. (II)
    $\partial \mathcal{C}/\partial W^l \leftarrow \boldsymbol{\Delta}^l \otimes \boldsymbol{a}^{l-1}$ ▷ eq. (IV)

6.4 Hidden Activation Functions

Hidden activation functions play an important role when it comes to the neural nets' ability to learn. For a long time Sigmoid functions was a common choice for σ. These functions are symmetric about the origin, with their characteristic S-shaped curve. The logistic function,

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

and the hyperbolic tangent are two popular Sigmoid functions commonly used as the activation function. A reason behind their popularity is their ability to produce outputs close to zero. They do however suffer the from the fact that $\sigma'(x) \approx 0$ for large x. Functions that exhibits this behaviour are so-called saturated activation functions. Since the derivative of $\sigma(x)$ plays an

important role in the backpropagation algorithm (see eq. (III)), this may cause a problem famously known as the the vanishing gradient problem. Not all activation functions are saturated, and two commonly introduced activation functions remedies this problem. The first one is the rectified linear unit (ReLU) which is given by

$$\sigma(x) = (x)^+ = \max(0, x). \tag{6.12}$$

The second activation function is known as the leaky ReLU, and is given by

$$\sigma(x) = \begin{cases} x & \text{if} \quad x > 0, \\ 0.01x & \text{else.} \end{cases} \tag{6.13}$$

Both of these avoid the problem with vanishing gradients, i. e., when $\sigma'(x) \to 0$ for large x, although they are not entirely without drawbacks. The ReLU activation function suffers from a problem commonly known as the dying ReLU, which is closely related to the vanishing gradient problem. It occurs when ReLU units only output 0 regardless of the input, and the unit becomes inactive. Lu et al. found that the problem could be alleviated by constructing deep and wide networks. However, this would highly affect the time usage and computational power needed for the training procedure [121].

6.5 Parameter Initialization

Initialization of weights has an important effect on the performance of the neural net. The fundamental problem is the weighted sum in eq. (6.2), which we repeat for convenience,

$$z_j^l = \sum_k W_{jk}^l a_k^{l-1} + b_j^l, \tag{6.2 revisited}$$

which can become very large (in absolute value) if the sum consists of many terms. This, combined with large weights, can further exacerbate the vanishing gradient problem. A solution to this problem is to decrease the likelihood that the sum becomes large. To robustly implement this, the weights can be randomly initialized according to $\mathcal{N}(0, 1)$ and further scaled with the squared root of the number of elements in the sum above. Let us loosely consider z_j^l as a distribution of possible values to be fed to the activation function. By interpreting it as a distribution, the scaling of the weights can be seen as reducing the width of this distribution, which reduces the likelihood of large (absolute) values of z.

The biases can be initialized according to $\mathcal{N}(0, 1)$ because they do not present the same scaling problem, i.e., they do not scale with the complexity of the model.

6.6 Regularization

There are fundamentally two ways to regularize a neural network: reducing the number of dimensions of the parameter space or reducing the effective size of each dimension [122]. This section will introduce one regularization method that temporarily reduces the number of parameters during training, namely dropout, and two methods that reduce the effective size of the dimensions of the parameter space, early stopping and weight decay.

6.6.1 Early Stopping

A simple yet effective regularization technique is early stopping. During training, it is normal that the model's performance is evaluated on a validation set. This evaluation estimates the out-of-sample error, which we want to keep as low as possible. A naive approach is to stop the

training procedure when the reduction of this error starts to stagnate or increase. Hence, we use the increase, or lack of decrease, in the out-of-sample error as a stopping criterion.

In the case where the out-of-sample error has one global minimum, determining the stopping criteria for when we terminate the training is trivial. However, in reality, this is rarely the case with the out-of-sample error having multiple local minima. This gives rise to the question of how we choose a reasonable stopping criteria when there is not a single global minimum. There exist several plausible ways to determine a cogent stopping criterion. Here we will describe two based on the work of Prechelt in [122].

The first stopping criterion involves a generalization metric that measures the relative increase of the out-of-sample error compared to the minimum error obtained so far in the training process. As mentioned in the section about optimization in chapter 4, a model is trained over a series of epochs. An epoch is a training cycle where all of the training data have been utilized in the optimization process. The in-sample and out-of-sample errors can then be describes as functions of epoch number t, $E_{\text{in}}(t)$ and $E_{\text{out}}(t)$. The minimum out-of-sample error up to epoch t can then be described as

$$E_{\text{opt}} = \min_{t' \leq t} E_{\text{out}}(t'). \tag{6.14}$$

The generalization metric is then defined as

$$G(t) = 100 \cdot \left(\frac{E_{\text{out}}(t)}{E_{\text{opt}}(t)} - 1 \right). \tag{6.15}$$

A high generalization value indicates overfitting, so one natural choice for a stopping criterion is to stop training when $G(t) > \alpha$ for some chosen threshold α.

Another approach is to stop training when the generalization value has increased in s successive steps. Here we assume that the ongoing increase indicates the beginning of overfitting. Since there is no guarantee that any of the two suggested stopping criteria will terminate the training, a general stopping criterion can be imposed as a safety net. If the out-of-sample error does not improve more than a chosen value β after a high number of epochs, cease training.

6.6.2 Weight Decay

Adding a L_2-regularization term to the cost function when working with neural nets is also called weight decay [123]. This can be achieved by defining the new cost-function

$$\mathcal{C}' = \mathcal{C} + \frac{\lambda}{2} \sum_l \left\| W^l \right\|_2^2, \tag{6.16}$$

where $\mathcal{C}$ is the un-regularized cost-function, λ is the regularization parameter and $\left\| W^l \right\|_2$ is the Frobenius norm of the matrix W^l. This matrix norm is defined as being

$$\left\| W^l \right\|_2 = \sqrt{\sum_i \sum_j |W_{ij}|^2}. \tag{6.17}$$

Here we are working with the squared norm of the weights due to computational efficacy. By squaring the norm, we remove the square root, simplifying the calculation of the gradient, which is also why the factor 2^{-1} appears in the regularization term. This adds an additional term to the update rule for the weights:

$$\frac{\partial \mathcal{C}'}{\partial W^l_{jk}} = \Delta^l_j a^{l-1}_k + \lambda W^l_{jk}. \tag{6.18}$$

If we then take one optimization step using gradient descent, the weights are updated with the following scheme

$$W^l_{kj} \leftarrow (1 - \eta_t \lambda) W^l_{jk} - \eta_t \frac{\partial \mathcal{C}}{\partial W^l_{jk}}. \tag{6.19}$$

Hence, the L_2-term has added a constant shrinkage factor to the weights in each optimization step.

6.6.3 Dropout

Dropout is a stochastic regularization technique that, during training, randomly drops selected nodes, including input nodes, first introduced by Hinton in 2012 [124]. To drop a node refers to the node being temporarily removed from the network with all its incoming and outgoing connections. The simplest case is that each node has an independent probability p, often set to be $p = 0.5$, of being dropped. For the input nodes, this probability is closer to 1 than to 0.5. This method mimics a sampling procedure where we sample a "thinned" network at each epoch or mini-batch, depending on the optimization algorithm. Hence, a neural net with a total of n nodes can be seen as a collection of 2^n thinned networks [125].

The complete network is used to test the model on unseen data. Then if a node had a dropout probability of p, the node is multiplied by p at test time. The regularization technique can thus be considered a form of model averaging, where we train our model on many different network architectures and combine them at test time to make more accurate predictions.

Chapter 7

Gradient Boosted Trees

Gradient boosted trees are an ensemble method, a class of methods that build on the same principle as dropout regularization for neural nets: they combine predictions from multiple statistical models to improve performance on prediction tasks. An ensemble describes a set of regressors (or classifiers) whose individual predictions are combined in some way, depending on the chosen method [126]. These regressors are often described as being "weak", simply referring to their performance being only slightly better than random guessing. The weak learners in gradient boosted trees are decision trees, which are the fundamental learning algorithm behind other tree-based models.

There are a number of different gradient boosted tree algorithms. For our methodological comparison in this thesis, we are interested in the Extreme Gradient Boosting algorithm, often abbreviated to XGBoost [127]. This algorithm incorporates ideas from gradient descent, the general concept of boosting and decision trees. This chapter is thus dedicated to presenting the underlying theory needed to understand the inner workings of this algorithm. Gradient descent was covered in section 4.5, so we start by giving a high-level introduction to boosting in section 7.2, followed by the theory behind decision trees. The final section of the chapter outlines the XGBoost algorithm. Before embarking on the journey described above, we briefly describe the intuitions behind ensembles and why they work.

7.1 Intuitions behind Ensembles

Dietterich [126] identified three fundamental reasons why ensembles sometimes work better than a single regressor. The first reason is statistical. As we recall from section 4.1, our goal when using a supervised learning algorithm is to explain the response through a mapping function $\hat{f}$ chosen from a hypothesis set $\mathcal{H}$. If the amount of training data is small compared to the hypothesis set, a statistical problem can arise where several mapping functions perform equally well on the training data. Instead of choosing one of the models and risking choosing the wrong one, we can combine all of the regressors into one ensemble, drastically lowering the chances of making the wrong pick.

Even though the amount of training data is sufficiently large, single learners can still encounter problems when fitting the model. The training process often includes some form of local search for the optimal model parameters, recall section 4.5. These searches can sometimes get stuck in local minima, giving a second reason why an ensemble can work better than an individual regressor. Diettrich describes this reason as being computational, remarking that performing local searches from many different starting points may approximate the true underlying function better than any of the individual models that simply start from a single point.

The final reason is described as being representational. This is related to the fact that in most cases, the true function cannot be represented by any of the functions in $\mathcal{H}$. Forming a

linear combination (or weighted sum) of several predictors from the hypothesis set may expand the space of representable functions. In principle, $\mathcal{H}$ is the set of all possible regressors, but given the finite nature of our training data, learning algorithms explore only a finite set of hypotheses.

7.2 High-Level Introduction to Boosting

As with any machine learning concept, there are many different boosting algorithms, but the majority of them are based on the same principles. The underlying idea is to learn several weak regressors iteratively and combine them in a weighted sum to build a strong aggregated boosted regressor. Denoting an ensemble of M weak regressors as $\{g_k(\boldsymbol{x})\}_{k=1}^{M}$, the aggregated regressor g_A can be expressed through the sum

$$g_A(\boldsymbol{x}) = \sum_{k=1}^{M} \alpha_k g_k(\boldsymbol{x}), \tag{7.1}$$

where α_k is the weight associated with regressor $g_k(\boldsymbol{x})$. All of the weights sum to one, i. e. $\sum_k \alpha_k = 1$. The weights indicate how much each regressor contributes to the aggregated regressor, and how each weight's value is assigned is dependent on the specific algorithm. Since we are adding regressors iteratively, each new regressor is constructed to be slightly improved compared to the previous one. How this behaviour is implemented varies from algorithm to algorithm, and we will give technical details of this procedure for XGBoost in section 7.4.

7.3 Decision Trees

Decision trees are a non-parametric supervised learning algorithm, and when used in regression tasks, they are often coined regression trees. In this section, we describe a popular method for tree-based regression called CART [88]. This method also encompasses the description of classification trees but is omitted in this discussion.

As before, denoting our dataset as $\mathcal{D}$, consisting of tuples on the form $\{(\boldsymbol{x}^{(i)}, y_i)\}_{i=1}^{N}$ with N observations and a single response. The feature vectors are p-dimensional real-valued vectors. The fundamental behaviour of the CART algorithm is to split the input data into two descendant subsets repeatedly. The splits are governed by a series of questions related to the data features, and the subsets that are not split into a new partition are called terminal nodes or leaves. The top node is referred to as the root node and is denoted t_0. In each terminal node t, a constant $\hat{y}_t$ is fitted, which correspond with the final predictions of the tree. These types of models are highly interpretable, and the feature space partition is fully described by a single tree [117].

When constructing a tree regressor, three key elements need to be determined: (i) a rule for assigning a constant value to every terminal node, (ii) a way to select a split at every non-terminal node and (iii) a rule for determining when a node is terminal and (iii) [88]. We start at the top and work our way down the key elements.

(i) Using the residual sum of squares as our quantitative measure for how much our predictions deviate from the truth, the constant that best minimizes this metric is the average of all responses y_i that end up in terminal node t. This can be easily proven by solving

$$\frac{\mathrm{d}}{\mathrm{d}a}\left(\sum_{i=1}^{N}(y_i - a)^2\right) = 0, \tag{7.2}$$

where a is a constant. Hence, the predicted value, $\hat{y}_t$, at node t is

$$\hat{y}_t = \frac{1}{N_t}\sum_{x_i \in t} y_i, \tag{7.3}$$

where N_t is the number of cases that end up in node t and the sum is over all y_i such that $x_i \in t$.

(ii) Next, we need to address how to split a non-terminal node. When discussing how to split a specific node, it will essentially be a terminal node since it has not yet been subjected to a split and will be denoted t. Hence, in the following discussion, a node will start as a terminal node t but end up as a non-terminal node. For every node t there is a candidate split s for a splitting variable j that creates two new nodes, t_L and t_R. These two nodes define the pair of half-planes

$$t_L(j,s) = \{\boldsymbol{x} | x_j \leq s\} \quad \text{and} \quad t_R(j,s) = \{\boldsymbol{x} | x_j > s\}. \tag{7.4}$$

The splitting variable j and split s is the pair that solve the minimization problem

$$\min_{j,s} \Big[\sum_{x_i \in t_L(j,s)} (y_i - \hat{y}_{t_L})^2 \quad + \quad \sum_{x_i \in t_R(j,s)} (y_i - \hat{y}_{t_R})^2 \Big]. \tag{7.5}$$

The variable j that best separates the high responses from the low ones will be the best splitting variable. The split point s can then be determined by scanning through every value for that variable and seeing which value reduces the error the most.

After performing an unknown number of splits, we are left with a set of splits and a set of terminal nodes, S and $\mathcal{T}$ respectively. The set of splits and their order determines the binary tree T with the terminal nodes $\mathcal{T}$. The complexity of a tree T is denoted $|T|$, and corresponds with the total number of terminal nodes in T.

(iii) The final element is then to establish a rule for when a node is terminal or not. The size of the tree is an important factor; a deep tree would easily overfit the data, while a shallow tree would have problems capturing the underlying structure of the data. Thus, the tree depth is a complexity parameter, a hyperparameter that must be tuned during training. Decision trees have the property of always experiencing a reduction (or no change) in their overall error for every split performed, so we need a clever stopping rule.

The strategy is to construct a large tree, denoted $T_{\max}$, and stop the splitting when the depth reaches a predefined stopping criterion. This criterion is usually when $N_t \leq N_{\min}$ for every $t \in \mathcal{T}$. The value of $N_{\min}$ is often taken as 5. We then prune $T_{\max}$ using a technique called minimal error-complexity pruning. Pruning is a general term describing the procedure where redundant partitions are removed from the tree, which reduces the complexity of $T_{\max}$.

7.3.1 Minimal Error-Complexity Pruning

A subtree of $T_{\max}$ is denoted T. Any tree that can be obtained by pruning $T_{\max}$ is a subtree, i. e. $T \subset T_{\max}$. Defining the error-complexity measure $R_\alpha(T)$ as

$$R_\alpha(T) = \frac{1}{N} \sum_{t \in T} \sum_{x_i \in t} (y_i - \hat{y}_t)^2 + \alpha|T| \equiv R(T) + \alpha|T|, \tag{7.6}$$

with $R(T)$ being the error of tree T and the second term imposes a penalty for its complexity, with $\alpha \geq 0$ being a complexity parameter. Hence, minimal error-complexity pruning is a form of shrinkage regularization. The complexity function can be described as a linear combination of the error of the tree and its complexity.

The idea is now to find a subtree $T_\alpha \subseteq T_{\max}$ that minimizes $R_\alpha(T)$ for each value of α. High values of α result in smaller trees, and for a sufficiently high value the pruned tree would only consist of the root node. If α is zero we are simply left with $T_{\max}$. For each α there is a unique subtree T_α that minimizes $R_{\alpha(T)}$, the proof is omitted but can be found in [88].

Even though the complexity parameter α is a continuous variable, there is a limited number of subtrees. The pruning process will produce a finite decreasing sequence of substrees,

$$T_1 > T_2 > \cdots > t_0,$$

with each subtree having successively fewer terminal nodes than the previous one, and an increasing sequence of the complexity parameter

$$0 = \alpha_1 < \alpha_2 < \cdots < \alpha_{\text{root}}.$$

The pruning process does not start from $T_{\max}$, but from a tree T_1, which is the smallest subtree of $T_{\max}$ that satisfy

$$R(T_1) = R(T_{\max}). \tag{7.7}$$

The fact that decision trees have the property of always experiencing a reduction or no change in its overall error for every split performed, i. e.

$$R(t) \geq R(t_L) + R(t_R), \tag{7.8}$$

means that there could exist a subtree without the two terminal nodes t_L and t_R such that eq. (7.8) is an equality. Pruning off the two redundant terminal nodes and repeating the process until a reduction in the error is reached yields the starting tree T_1. Searching through the whole sequence of subtrees to obtain the minimizing subtree T_α is computationally expensive. Instead, it is found from weakest link pruning which we will describe next.

7.3.2 Weakest Link Pruning

Any internal node can be interpreted as a root node for a smaller branch of a tree. Denoting a branch rooted at internal node t as T_t, the error for this branch is then $R(T_t)$. The complexity error $R_\alpha(T)$ can also be calculated for a single node or branch. For $t \in T_1$,

$$R_\alpha(t) = R(t) + \alpha, \tag{7.9}$$

and for branch T_t,

$$R(T_t) = R(T_t) + \alpha|T_t|. \tag{7.10}$$

When $\alpha = 0$, $R_0(T_t) < R_0(t)$ since the branch T_t has a finer partitioning of the data than the branch root t. When increasing α, $R_\alpha(T_t)$ will increase more rapidly than $R_\alpha(t)$ since $|T_t| > 1$. So for a specific value α',

$$R_{\alpha'}(T_t) = R_{\alpha'}(t). \tag{7.11}$$

When $\alpha > \alpha'$, the error in the branch will be higher than the branch root. The internal node in T_1 that achieves the equality in eq. (7.11) for the lowest value of α' is called the weakest link, and the branch belonging to that node is removed.

The procedure above can be systematized by defining a function $g_i(t)$ for every i in the sequence of subtrees,

$$g_i(t) = \begin{cases} \dfrac{R(t) - R(T_t)}{|T_t| - 1}, & t \in T_i, t \notin \mathcal{T}_i \\ +\infty, & t \in \mathcal{T}_i. \end{cases} \tag{7.12}$$

The fraction is obtained by solving the inequality $R_\alpha(T_t) < R_\alpha(t)$ for α, and it describes the ratio between the differences in error and complexity between the branch and branch root. The weakest link $\bar{t}_i$ is then

$$\bar{t}_i = \underset{t \in T_i}{\arg\min}\, g_i(t), \tag{7.13}$$

and its belonging branch is pruned by forming the new tree

$$T_{i+1} = T_i - T_{\bar{t}_i}. \tag{7.14}$$

If there are multiple weak links, all of the branches belonging to these links are also removed. We gradually increase α by setting $\alpha_{i+1} = g_i(\bar{t}_i)$. This iterative process starts from $i = 2$, after obtaining T_1 as described above with $\alpha_1 = 0$, and continues until we have produced the root node as the final subtree. This sequence must then contain T_α.

The final step is to estimate α using five or 10-fold cross-validation. The value of α that minimizes the residual sum of squares the most is then chosen as our estimated value, $\hat{\alpha}$, and our final tree is then $T_{\hat{\alpha}}$.

7.4 Extreme Gradient Boosting

The extreme gradient boosting (XGBoost) algorithm boosts ensembles of CARTs and has become one of the most widely used machine learning techniques for a variety of supervised learning tasks. It is a versatile algorithm, and in 2015, out of the 29 winning solutions published at Kaggle[1], 17 used XGBoost [127].

7.4.1 Parametrization of CARTs and Cost Function

Denoting the kth tree as $g_k(\boldsymbol{x})$, using the weak regressor notation introduced in section 7.2, with $|T|$ number of terminal nodes. The tree is parametrized using two quantities: a function $q(\boldsymbol{x})$ that maps the input vector to one of the terminal nodes, and a weight vector $\boldsymbol{w} = (w_1, \ldots, w_{|T|}) \in \mathbb{R}^{|T|}$ that stores the constant prediction value of each terminal node. The function q also represents the structure of the tree. The prediction of tree $g_k(\boldsymbol{x})$ can then be expressed as $w_{q(\boldsymbol{x})}$. Since we are working with ensembles, we explicitly express the prediction of several CARTs for a datapoint $(\boldsymbol{x}^{(i)}, y_i)$ as

$$\hat{y}_i = g_A(\boldsymbol{x}^{(i)}) = \sum_{k=1}^{M} g_k(\boldsymbol{x}^{(i)}), \qquad g_k \in \mathcal{F} \tag{7.15}$$

where M is the total number of trees in the ensemble and $\mathcal{F}$ is the set of all possible CARTs. To be able to fit our model in eq. (7.15) we need to define a cost function. As mentioned in the section about decision trees, CARTs will almost always experience a reduction in its error when performing a split, so the cost function must include a regularization term. The cost function in XGBoost is

$$\mathcal{C}(X, g_A) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{j=1}^{M} \Omega(g_j), \tag{7.16}$$

where l is a differentiable and convex function and Ω is the regularization function,

$$\Omega(g) = \gamma|T| + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2. \tag{7.17}$$

Here γ and λ are regularization hyperparameters. The regularization function penalizes both large weights on the terminal nodes and the depth of the tree. For regression tasks, l is often the mean squared error.

7.4.2 Optimization

Given that the cost function in eq. (7.16) is a composite function, it cannot be optimized using descendants of gradient descent or other traditional optimization methods in Euclidean space

[1]Kaggle is an online community for data scientists that regularly hold machine learning competitions.

[127]. Instead the model has to be trained in an additive fashion, meaning that we iteratively add the tree to our ensemble that most reduces the error in eq. (7.16). Denoting the prediction at iteration t of the ith instance as $\hat{y}_i^{(t)}$, the function we need to minimize at iteration step t is

$$\mathcal{C}^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + g_t(\boldsymbol{x}^{(i)})\right) + \Omega(g_t). \tag{7.18}$$

The idea is that each tree we add is a small perturbation of the previous tree, and hence we can approximate the loss with a second order Taylor expansion. The reason for making this expansion is that if l is not the mean squared error, but some other loss function, obtaining the total loss may not be as straight forward. The approximated loss is

$$\mathcal{C}^{(t)} \approx \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t-1)}) + u_i g_t(\boldsymbol{x}^{(i)}) + \frac{1}{2} v_i g_t(\boldsymbol{x}^{(i)})^2 + \Omega(g_t), \tag{7.19}$$

with u_i and v_i being

$$u_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \qquad \text{and} \qquad v_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}). \tag{7.20}$$

Here the notation $\partial_x^n f$ refers to the nth partial derivative of f with respect to x. We are interested in adding a tree that reduces the error in eq. (7.19). The first term in this equation is constant with respect to g_t, removing it yields the simplified cost $\tilde{\mathcal{C}}^{(t)}$ at iteration t,

$$\tilde{\mathcal{C}}^{(t)} = \sum_{i=1}^{N} \left[u_i g_t(\boldsymbol{x}^{(i)}) + \frac{1}{2} v_i g_t^2(\boldsymbol{x}^{(i)}) \right] + \Omega(g_t). \tag{7.21}$$

The next step is to determine the structure of the tree g_t that yields the highest reduction of the simplified error. Next, we derive the optimal weights and loss for a tree g_t that minimizes eq. (7.21) and use this information to determine the structure of the new tree.

Defining the set of input vectors $\boldsymbol{x}^{(i)}$ that map to terminal node j as the instance set $I_j = \{i | q(\boldsymbol{x}_i^{(i)}) = j\}$, the simplified cost can be recast into the following expression

$$\tilde{\mathcal{C}}^{(t)} = \sum_{j=1}^{|T|} \left[\sum_{i \in I_j} u_i w_j + \frac{1}{2} \left(\sum_{i \in I_j} v_i + \lambda \right) w_j^2 \right] + \gamma |T|. \tag{7.22}$$

Remembering that the prediction of tree $g_j(\boldsymbol{x})$ is equal to $w_{q(\boldsymbol{x})}$, and inserting this into eq. (7.21), along with the explicit expression for the regularization function Ω, results in eq. (7.22). Holding the tree structure fixed, the optimal weight w_j^{opt} of terminal node j is

$$w_j^{\text{opt}} = -\frac{\sum_{i \in I_j} u_i}{\sum_{i \in I_j} v_i + \lambda}. \tag{7.23}$$

The corresponding optimal value of the cost is

$$\tilde{\mathcal{C}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{|T|} \frac{\left(\sum_{i \in I_j} u_i\right)^2}{\sum_{i \in I_j} v_i + \lambda} + \gamma |T|. \tag{7.24}$$

The expression for the optimal cost is sometimes called the structure score and is used to measure how good the quality of a tree structure q is. With the ability to quantify the goodness of the structure, we are now ready to learn the new tree.

In an ideal world, we would enumerate all possible tree structures and choose the one that best minimizes eq. (7.4.2). This is intractable, so instead, XGBoost constructs a tree from one split at a time through what is known as an approximate greedy algorithm. Greedy algorithms

do not necessarily find the optimal solution but can find local minima, which suffices in many cases. The approximate algorithm proposes several splitting points based on the percentiles of the feature distributions in the case where the features are continuous. The split candidates are then evaluated through what is known as the gain score. The gain is defined as the structure score of the new left node, the new right node and the original node:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} u_i)^2}{\sum_{i\in I_L} v_i + \lambda} + \frac{(\sum_{i\in I_R} u_i)^2}{\sum_{i\in I_R} v_i + \lambda} - \frac{(\sum_{i\in I_L} u_i + \sum_{i\in I_R} u_i)^2}{\sum_{i\in I_L} v_i + \sum_{i\in I_R} v_i + \lambda}\right] - \gamma, \tag{7.25}$$

where the subscripts L and R corresponds with the left and right terminal node, respectively. The regularization parameter γ belongs to the additional terminal node. The feature and split point that yields the highest gain are kept in the tree. The gain score also indicates if adding a feature to a branch improves its overall predictive ability. A feature with a higher gain can thus be said to have a higher impact on the predictions compared to features with lower scores.

7.4.3 Hyperparameters

The XGBoost algorithm has a number of hyperparameters that can highly affect the final result, and there are mainly three types of parameters: general-, booster- and task parameters [128]. The general parameters are related to which booster we are using and more general settings like the number of threads used during training. In our case, we will simply be using the default booster, tree ensembles. This section is thus dedicated to giving a short introduction to some of the most important booster- and task parameters and how they influence the training process. The values of these parameters are important for reproducibility purposes, and the parameters listed below are obtained through a tuning process. For the complete list, the reader is referred to the code documentation of the XGBoost package in [128].

Parameters for Tree Booster

`num_boost_round` The number of trees added to the ensemble is governed by this parameter. The parameter can be set during training, meaning that we do not need to fix it initially. Since trees are added iteratively to the ensemble, XGBoost tests our model at each iteration and checks whether a new tree improves our performance. This hyperparameter is the same as `n_estimators` used in the scikit-learn application programming interface (API) for XGBoost.

`max_depth` This parameter sets the maximum depth of a tree. A high value corresponds to a more complex model, which increases the model's chances of overfitting. The value of this parameter corresponds with the number of acceptable nodes from the root node to a terminal node.

`eta` Eta is a regularization parameter between 0 and 1 and is also called the learning rate or shrinkage factor. After the optimal weights $\boldsymbol{w}^{\text{opt}}$ of a new tree are obtained, they are scaled with `eta`, i. e., `eta` shrinks the new weights. This makes the training process a bit slower and the need for a higher number of trees added to the ensemble. It has been found that regularization with shrinkage provides superior results compared to cases where the number of components in the ensemble has been limited in order to prevent overfitting [129].

`min_child_weight` This parameter determines how many samples are needed in one node in order to keep partitioning the data. If the number of samples in a node is below this threshold, the node becomes a terminal node. This parameter also reduces the chances of the model overfitting.

`subsample` To prevent overfitting, we can construct trees using subsamples of the whole dataset, such that each tree is constructed on a slightly different dataset at each iteration. The `subsample` parameter is the ratio of training data used in each boosting round and can be set to values in the interval $(0, 1]$.

`colsample_bytree` This parameter describes the subsampling ratio of features used in each tree. As with `subsample`, this parameter have a range of possible values in the interval $(0, 1]$ and prevents overfitting.

Learning Task Parameters

`objective` This parameter describes the loss function and the underlying learning task. For regression `objective` is set to `reg:squaredloss`, referring to the learning task being regression and the cost function being the squared error, $(y_i - \hat{y}_i)^2$.

`eval_metric` Metric for evaluating the validation data during training. The default value for regression tasks is the RMSE metric.

Chapter 8

Methods

The focus of this thesis is formulated into three research aims. We presented them in the introduction, but repeat them here for convenience:

1. Investigate the ability of machine learning methods to predict a continuous measure of symptoms of anxiety and depression in new mothers, using data from a large population-representative prospective cohort.

    1.1. Predict levels of anxiety and depression at 6, 18 and 36 months postpartum using prenatal exposures measured at 17 and 30 weeks of pregnancy.

    1.2. Predict levels of anxiety and depression using exposures measured concurrently at 6, 18 and 36 months postpartum.

    1.3. Investigate the performance of modern machine learning approaches to statistical methods traditionally employed in psychology, such as linear regression.

    1.4. Compare the performance of models using i) aggregate scores on established scales, ii) item-level analyses, iii) dimensional reduction by principal component analysis, and iv) data without dimensional reduction.

    1.5. Evaluate the performance of different methods for identifying individuals at risk for clinical levels of depression and anxiety.

2. Use variable importance methods to investigate whether machine learning methods can contribute to identifying prenatal or concurrent exposures that can be of clinical interest or help inform theoretical models of post-party emotional problems.

    2.1. Identify prenatal exposures associated with increased risk of symptoms of anxiety or depression at 6, 18 and 36 months after birth.

    2.2. Determine whether different sets of variables best predict internalizing symptoms at 6, 18 and 36 months in the extended postpartum period.

    2.3. Compare the sets of variables identified through traditional linear model methods and machine learning approaches.

3. Present a set of recommendations on using machine learning to analyze registry and health survey data based on the experiences from answering the first and second research aims.

This chapter presents the sample base for our numerical experiments, measures and instruments included in the sample, and a description of the numerical experiments and how they relate to the research aims. We describe how they were conducted following the framework outlined in chapter 4. All of the numerical work related to this thesis was performed using Python, and a small section about code availability can be found at the end of this chapter.

8.1 Sample: The Norwegian Mother, Father and Child Cohort Study

The Norwegian Mother, Father and Child Cohort study (Moba) is a study aimed to discover new knowledge about the causes of disease and health issues among mothers and children [42]. Again, see Appendix C for a description of cohort studies. The study invited 277 702 women to participate between 1999 and 2008, with an acceptance rate of 41.0% [33]. The women were recruited during a routine ultrasound examination that all Norwegian women are offered around week 17 after gestation. A postal invitation was later sent out, which, among other things, included an informed consent form and an information brochure. There were no exclusion criteria for participating in the study. However, the study was conducted in Norwegian and thus excluded non-Norwegian-speaking women. No compensation was given for participating in the study. The original goal was to recruit women from 100 000 pregnancies. In addition to the mothers, over 75 000 fathers were also recruited to participate in the study [33].

8.1.1 Data Collection

The data was collected by administering several self-reporting questionnaires to the mothers and fathers through the mail over an extended period. A woman can participate in the study with more than one pregnancy, making the number of pregnancies greater than the number of participating mothers, respectively 112 645 and 95 136. Hence, the unit of observation is each pregnancy and not the individual mother [130]. Each mother, father, and pregnancy has a unique id in the data, making it possible to monitor each pregnancy's trajectory between questionnaires and identify several pregnancies linked to the same parents.

The first questionnaire (Q1) was sent out during the 17th week of pregnancy, and it focused on the mother's mental and physical health prior to and during the pregnancy. Next, a questionnaire (Q2) regarding the mother's diet during the pregnancy was administered at week 22. A third questionnaire (Q3) followed up on the mental and physical health of the mother after 30 weeks of pregnancy.

After delivery, the mothers were followed up with questionnaires when the child was six, 18 and 36 months old, in questionnaires coined Q4, Q5 and Q6, respectively. All of the surveys focused on maternal health and the child's development. At ages five, seven and eight, the main topics of the questionnaires were the child's health, lifestyle and mental development. These questionnaires are logically named Q-5y, Q-7y and Q-8y. The three latter questionnaires were first introduced in 2010, meaning that several of the participating mothers did not receive these forms. Based on our research aims, our analyses only to containing data from the six first questionnaires. A timeline of the relevant data-collection time points is shown in Figure 8.1.

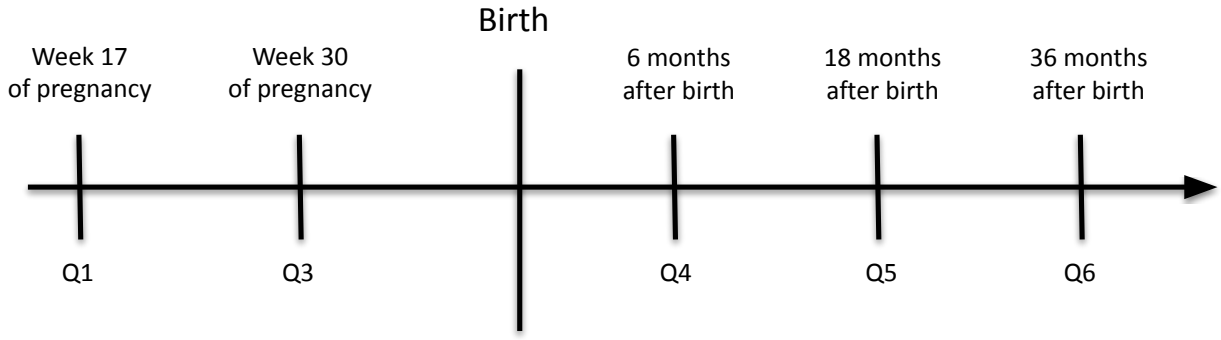

Figure 8.1: The timeline showing when the specific questionnaires were answered by participants in the MoBa study. The figure is adapted from [2].

In addition to the self-reporting questionnaires, the study also collected blood and urine samples during the first ultrasound appointment. This thesis focuses on the analysis of the answers to the questionnaires. For a more detailed description of the content in each of the specific questionnaires, we refer the reader to their instrument documentation found in [131].

Due to an extensive data collection period, various changes have been made to the questionnaires over the years. Questionnaires Q1 and Q6 come in four different versions, while Q3, Q4 and Q5 have five versions. The different versions are encoded with the first letters of the alphabet. The majority of the changes between the versions are minor, with the most noticeable differences being between version A and later versions. The discrepancies are frequently connected to the formulation of an item or having added or removed items between versions.

The response rates for Q1 and Q3-Q5 were 95.1%, 91.4%, 87.0% and 61.4%, respectively. In a failed attempt to increase the response rate, questionnaires administered at weeks 22 and 30 during the pregnancy were made electronic. Only 17% answered the online version of Q2, and 13% for Q3 [130].

Given the longitudinal nature of the study, some topics are repeatedly covered in separate questionnaires. Each question has a unique question ID, meaning that if the same question is asked at two different time points, there are two question IDs. Hence, the total number of features in all of the questionnaires combined exceeds 7000.

8.2 Measures

This section outlines the different measures and instruments included in the different self-reporting questionnaires.

Depression and Anxiety

Our outcome variable is the level of depression and anxiety symptoms. They are measured using two short versions of the 25-item version of the Hopkins Symptoms Checklist described in chapter 2. The SCL-5 was included in Q1, and it is estimated to correlate 0.92 with the total score from the original instrument. For questionnaires Q3-Q6, the SCL-8 was included, and it has a 0.94 correlation with the SCL-25 [132, 133]. The SCL-8 instrument encompasses SCL-5, and the items are listed in Table 8.1 with the respective response options. Three of the questions in SCL-5 and four out of eight questions in SCL-8 measured depressive symptoms, while the remaining measured anxiety symptoms.

Table 8.1: The table lists all eight items in SCL-8, along with their response options. The first five questions constitute SCL-5.

Have you been bothered by any of the following during the last two weeks?	Response Options
1. Feeling fearful	1-Not bothered 2-A little bothered 3-Quite bothered 4-Very bothered
2. Nervousness or shakiness inside	
3. Feeling hopeless about the future	
4. Feeling blue	
5. Worrying too much about things	
6. Feeling everything is an effort	
7. Feeling tense or keyed up	
8. Suddenly scared for no reason	

All items are scored using a Likert scale [55] from 1 to 4, with 4 being the highest indicator of

depressive or anxiety symptoms. For all versions of the symptom checklist, the total SCL score is the mean of all responses. A mean score greater or equal to 1.75 is considered to be a strong predictor of a mental disorder in need of treatment in the SCL-25 [56]. Several cut-off points have been established for the Norwegian population. For SCL-10, the cut-off is 1.85, while for SCL-5, it is 2.0 [57]. The mean SCL score, a continuous variable on the interval $[1, 4]$, will be our dependent variable in the numerical analyses. The distributions of the target variable at the different time points are given in Figure 8.2.

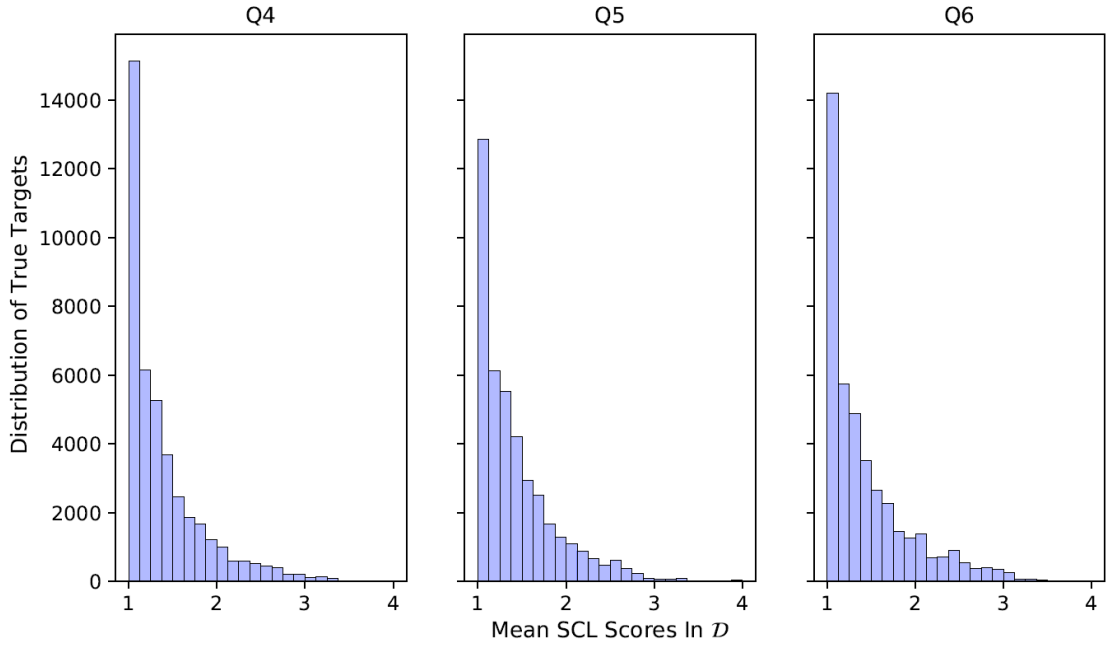

Figure 8.2: The distribution of the target variable, the mean SCL score, at the time points Q4, Q5 and Q6. From the figure, it is evident that the target variable is right-skewed, with a majority of the participants having a mean score below 2.

Mother's Mental Health

Several items and instruments measure different aspects of the mother's mental health. Postnatal depression was assessed six months after delivery in Q4 through the Edinburgh Postnatal Depression Scale (EPDS) described in chapter 2. Past history of depression was assessed in Q1 with the scale lifetime history of major depression [134], consisting of eight single items. Seven out of the eight items had a binary response option, while the last one was encoded on a ratio scale. The first six items were related to mood and behavioral problems, and the seventh item asked if the participant had had three or more of these problems simultaneously and for how many weeks the longest period lasted. The final item asks if there was any particular reason for having these problems. The items are closely related to the DSM-III criteria for lifetime major depression [132].

Questions about weight control and eating disorders were included in Q1, Q5 and Q6. The questions related to the eating disorders were designed following the DSM-IV diagnoses of anorexia nervosa, bulimia nervosa and eating disorders not otherwise specified [132]. In the sixth questionnaire, an instrument for adult ADHD was included, the adult ADHD self-report scale [135]. The scale consists of six items, and it is a short version of an 18 items scale. The responses are encoded on a Likert scale from 1 to 5.

Socio-Demographic Variables

Several items measured socio-demographic factors through education- and income level, family household, living situation and immigration status. The socio-demographic variables were assessed in the first questionnaire, while some re-occurred in later questionnaires.

The level of education of the parents was measured with four items. The two first items contained information about the highest degree of completed education for the mother and father, while the two remaining items asked if the parents were currently pursuing a higher education degree. Income level was assessed with three items related to the parents' yearly gross income and if the household could financially manage without the mother's income. The living situation and family household were measured by asking how many people the mother lived with and their type of residence. In Q5 and Q6, items related to the child's living situation were included, explicitly measuring if it lived with its father or not. The financial situation was again assessed in Q4 with one item. Immigration status was implicit accounted for with items asking about the parents' native language.

Personality

Several instruments can be linked to different aspects of personality. General self-efficacy (GSE) was measured in Q3 and Q5 using a 5-item version of the 10-item GSE Scale by Schwarzer and Jerusalem [136]. The scale uses a 4-point Likert scale from "not at all true" (1) to "exactly true" (4) regarding statements about problem-solving. A scale named the differential emotional scale by Izard et al. [137] was included in every questionnaire, except the first one, to capture how often a participant experienced enjoyment and anger in her everyday life. Six items from the scale were included, and it was administered on a 5-point Likert scale.

All five questionnaires included four items from Rosenberg's self-esteem 10-item scale [138]. The items were related to how the participant felt about herself, and it used a 4-point Likert scale.

Interpersonal Relationships

Interpersonal relationships accounted for in the self-reporting questionnaires include the relationship satisfaction between the mother and her partner, civil status and general social relations and support. The relationship satisfaction was measured in all five questionnaires using the relationship satisfaction scale [132], consisting of ten items explicitly developed for MoBa. The responses to the items are encoded using a Likert scale from 1 to 6. Civil status was measured with one item in all questionnaires except Q4. Social relations and general social support were measured in three items in both Q1 and Q3.

The Mother's Well-Being

A variety of items and scales measured the mother's well-being and other aspects that could influence the general well-being. One such instrument was the satisfaction with life scale (SWLS) by Diener et al. [139]. The scale is a 5-item instrument designed to measure global cognitive judgments of satisfaction with one's life. All answers are scored on a 7-point Likert scale from "strongly disagree" (1) to "strongly agree" (7). The SWLS scale was included in all questionnaires except Q5. In the fifth questionnaire, the short version of the world health organization's quality of life instrument [140] was included. This short version included a number of items measured on different Likert scales.

Items related to the experience of abuse and assaults were included in Q1, Q3 and Q6. The topic received a more extensive covering in Q3, with, in total, four questions about emotional-, physical- and sexual abuse. To each question, there were three responses: i) no, never, ii) yes, as

a child and iii) yes, as an adult. There were two additional questions about who the perpetrator was and if it happened during the last year for each question.

Participants were also asked about the experience of adverse life events in four out of the five questionnaires, with Q1 being the exception. The questions addressed if the participants had experienced any adverse life events since answering the last questionnaire, which explains why it was not measured in the first time point. There are in total 11 questions developed specifically for MoBa, each with a binary response option.

Employment and Absence from Work

Several items were included in all questionnaires to assess the mother's general employment situation and absence from work. These items measured a usual number of paid working hours, general absence from work, the parents' working situation (e.g., self-employment, employed in the private sector) and occupation title. Sick leave between the different questionnaires was measured from Q3 to Q5. The first and third questionnaires also assessed if the mother experienced any strains related to her work, such as a stressful environment, heavy lifting and monotonous tasks.

Substance Use

The participants were asked about their use of alcohol, illegal drugs and tobacco in the different questionnaires. Hazardous drinking behaviour was assessed in all questionnaires, either with single questions about alcohol intake or items adapted from two screening tools designed to measure harmful alcohol consumption: the alcohol use disorder identification test [141] and Rutgers alcohol problem index [142]. The first instrument had a mix of binary items and questions encoded on a Likert scale, while the latter had answers given on a Likert scale from 1 to 3.

The use of illegal substances, such as drugs and anabolic steroids, was measured in the three first questionnaires (Q1, Q3 and Q4) with binary response variables. Smoking and the use of tobacco were measured in all questionnaires.

The Child's Development and Behaviour

In the questionnaires following childbirth (Q4-Q6), there are several items and instruments related to the child's development and behaviour

The child's development was measured using the Ages and Stages Questionnaires (ASQ) [143]. The ASQ is developed for specific ages, so the specific version corresponding to the child's age in the different questionnaires was included. The items are encoded on a Likert scale from 1 to 3. In Q4, the child's mood and temperament were assessed using the Infant Characteristics Questionnaire – 6 months form [144], consisting of 11 items in total. In the fifth and sixth questionnaires, the temperament was measured using the emotionality, activity and shyness temperament questionnaire [145], which consisted of 13 items encoded on a Likert scale from 1 to 5.

The child's behaviour was assessed by the child behaviour checklist by Achenbach [146] in both Q5 and Q6. There are two versions of the checklist, the preschool- and the school-age checklist. The two versions have 100 and 120 items, respectively, and are all encoded on a Likert scale from 1 to 3.

In the fifth questionnaire, two screening instruments for autistic traits, ESAT [147] and M-CHAT [148] were included. They both consist of a series of items with binary response options. The ESAT instrument measures early social-communication skills, play and restricted and repetitive behaviours. The M-CHAT checklist includes items related to language, sensory

responsiveness, motor functions and social and emotional functions. It is advised that the two screening tools should be looked at together [149].

The Child's Communication and Social Skills

The questionnaires from the two latest time points, when the child was 18 and 36 months old, included various items and instruments measuring the child's communication skills. The two time points included an autism screening tool that measures non-verbal communication. The scale was named the non-verbal communication checklist [150] and consisted of four items encoded on a Likert scale from 1 to 3. In the final questionnaire, Q6, social communication was assessed with the social communication questionnaire [151], which is a parental-report autism screening tool consisting of a number of binary items. Social skills were measured through the strength and difficulties questionnaire by Goodman [152], which consists of six items measured on a Likert scale from 1 to 3.

The Child's Everyday Life and Well-Being

A variety of items assessed aspects of the child's everyday life and well-being. In Q5 and Q6, an item measured the amount of time the child spent in front of the TV another how much time the child spent outside. The two latest questionnaires included several items related to the child's daycare situation.

Other Measures

Several other items were also included in the different questionnaires. In Q1, previous pregnancies, and complications related to these, are assessed with a number of items. The sixth questionnaire contains items related to how the child control parts of the parents' life and how the parents perceive their ability to control their child's behaviour. There are, in total, nine items administered on a 9-point Likert scale.

Missing Items

As stated above, several items were missing from the available data. From the first questionnaire, items addressing previous and current illnesses and health problems, the use of contraception, menstruation, the use of medication and other supplements, different types of exposure, e.g., to radiation, harmful substances and noise, smoking, physical activity and problems with urination are missing.

Missing items in the third questionnaire were related to antenatal care and check-ups, hospitalization, health problems and incontinence during the pregnancy, medication and physical activity, food and beverage consumption, feelings related to childbirth, smoking and different exposures. In Q4, the missing items were related to the child's nutrition and sleeping habits, illness and health problems with the mother and the child, and medication use.

The missing items in the fifth and sixth questionnaires were related to the child's food and drink consumption, allergies, illnesses, health problems and medication use of the mother and child, the child's sleeping pattern, maternal concern, and physical activity.

8.3 Statistical Analyses

8.3.1 Numerical Experiments

Here we describe the two experiments created to reach and fulfill all of our research aims. We have added footnotes in the description to highlight how we incorporated the different sub-aims into the experiments. A section explaining the theoretical motivation behind a selection

of choices made when constructing the numerical experiments is included after the experiment descriptions.

Experiment 1: Investigating Predictive Ability and Feature Importance using Prenatal Exposures

The first numerical experiment revolved around investigating how the different methods were able to predict levels of depression and anxiety in the extended postpartum period when the training data consisted of different numbers of independent variables[1] collected during the prenatal period[2]. We created four different datasets from the items in Q1 and Q3. We trained multiple linear regression models, elastic nets, neural networks and gradient boosted regression trees on all four datasets for three different time points[3], and predicted the mean SCL score at each time point.

Out of the four datasets, three were smaller subsets, and the fourth included all of the independent variables from the prenatal period. The two first subsets were created based on a theoretical approach, where domain-specific knowledge presented in chapter 2 was applied to identify confounding features, and the third was created through a principal component analysis. The process is more thoroughly described in the following section.

In order to use the results from the experiment to acquire new knowledge about depression and anxiety in the extended postpartum period[4], we relied on the interpretability of the models' predictions. We only investigated the predictions from two of the datasets created, the dataset consisting of the principal components and the one including all available items. For the linear regression model and the elastic net, the sign and magnitude of the regression coefficients were inspected, while for the gradient boosted regression trees, the gain score was of interest[5].

When investigating the predictions from the dataset containing all available features, we clustered the single items into smaller groups before calculating the mean contribution of that group. An item was placed in a cluster based on which instrument the single item belonged to. The mean contribution of a group refers to the mean of either the regression coefficients or the gain scores of all items belonging to one group. To learn something from the predictions made by the principal components, we looked at the 20 components with the highest coefficients and gain scores and identified the top 20 single items with the largest weights in the linear combination. The 400 items were then grouped, and how often a group appeared was used as a measure of importance.

We note that the number of items in one group could vary, and some groups consisted of only one or two items. However, since we were looking at the 20 largest weights in the top 20 principal components, the number of times a single item could be counted was 20, making it possible for smaller groups to make a noticeable impact.

[1]Sub-aim 1.4: Compare the performance of models using i) aggregate scores on established scales, ii) item-level analyses, iii) dimensional reduction by principal component analysis, and iv) data without dimensional reduction.

[2]Sub-aim 1.1: Predict levels of anxiety and depression at 6, 18 and 36 months postpartum using prenatal exposures measured at 17 and 30 weeks of pregnancy.

[3]Sum-aim 1.3: Investigate the performance of modern machine learning approaches to statistical methods traditionally employed in psychology, such as linear regression.

[4]Sub-aim 2.1: Identify prenatal exposures associated with increased risk of symptoms of anxiety or depression at 6, 18 and 36 months after birth.

[5]Sub-aim 2.3: Compare the sets of variables identified through traditional linear model methods and machine learning approaches.

Experiment 2: Investigating Predictive Ability and Feature Importance using Concurrent Exposures

The second experiment used data collected at 6, 18 and 36 months postpartum[6], to train multiple linear regression models, elastic nets, neural networks and gradient boosted regression trees for prediction purposes[3] Data from each specific time point constituted a separate dataset, and we predicted the mean SCL score at each time point.

As in the first experiment, we compared the feature importance measures from the multiple linear regression models, elastic nets and gradient boosted regression trees. However, the comparison was made between the different models at all time points to determine whether different sets of variables best predict internalizing symptoms at 6, 18 and 36 months in the extended postpartum period[7,5]. We clustered the items in the same way described in the preceding section. However, we here ranked the different importance measures, meaning that we sorted the absolute value of them, to be able to compare the various measures at each time point.

Theoretical Motivation

Regression versus Classification Throughout this thesis, we have kept our focus solely on regression. In the literature, however, classification seems to be the favored statistical tool for researching postpartum depression and anxiety, with several studies framing their problems as a binary classification problem [17–20]. The assessment of postpartum depression and anxiety in these studies is often completed by either the SCL or EPDS instruments. The binary outcome variables are constructed from cut-off scores. There are several strategies for determining an appropriate cut-off score, such as using different criteria in a receiver operating characteristic (ROC) curve analysis [153]. The cut-off determines the instrument's sensitivity and specificity, with sensitivity being the likelihood of correctly identifying a diseased person when the instrument indicates that a disease is present, while specificity is the probability of not detecting a disease when the participant is healthy. More specifically, the two measures are given by

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{Total number of sick individuals}}, \tag{8.1}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{Total number of healthy individuals}}. \tag{8.2}$$

The two measures find themselves in a tradeoff, much like the bias and variance of a model, and increasing the sensitivity will lead to a decrease in specificity and vice versa.

A widely used criterion for determining the cut-off point is the point where the instrument sensitivity equals the instrument specificity [153]. However, some cases require a more sensitive instrument, e.g., when there are serious complications related to a disease. Sometimes, a higher specificity is needed if subsequent testing and diagnostic are expensive. In a Norwegian population, the cut-off scores for SCL-10 corresponded with an 89% sensitivity and 98% specificity, while the cut-off for SCL-5 had an 82% sensitivity and 96% specificity [57]. Given our interest in the tail of the distribution in Figure 8.2, high sensitivity is preferred over high specificity. We circumvent the possible pitfalls associated with the sensitivity-specificity tradeoff by not converting our problem into a binary classification problem. We also retain more statistical power by not thresholding the scores.

[6]Sub-aim 1.2: Predict levels of anxiety and depression using exposures measured concurrently at 6, 18 and 36 months postpartum.

[7]Sub-aim 2.2: Determine whether different sets of variables best predict internalizing symptoms at 6, 18 and 36 months in the extended postpartum period.

Another issue that arises with cut-off scores is correctly evaluating the values that lie in the area around the cut-off point. Differentiating between individuals with similar scores on different sides of the cut-off, without further interviews or supporting instruments, can lead to misclassification of individuals. Oftentimes, an analysis has some higher purpose beyond achieving a high predictive ability, e.g., to be used as a screening or diagnostic tool. It is important to note that we do not argue that we bypass the need for further assessments in a diagnostic process by using regression.

Apart from the challenges associated with categorizing a continuous target, we gain an increased statistical power by keeping the outcome variable in its original form. As most mothers do not experience high levels of depression and anxiety (ref. Figure 8.2), we can investigate increases in symptom levels in the non-clinical range.

Predictive Ability on Different Datasets Research aim 1.4 focuses on comparing how the predictive ability changes when we use different procedures for reducing dimensionality. A more general description of why reducing the dimensionality in the first place can be fruitful was given in section 4.2.

The use of aggregated variables is a commonly used technique in psychological research and is also used in several studies performed on the MoBa dataset [20, 35, 47]. The idea behind aggregation is that a set of multiple observations on a number of features describing the same underlying behaviour or phenomenon is more stable than any single observation [154]. By combining several features, possible measurement errors in specific features can be averaged out, resulting in a more reliable feature. However, the risk of losing information in the aggregation process is always present when combining multiple features. By comparing the predictive abilities of two datasets, one with aggregated features and one with the single items constituting the aggregated variables, we are able to determine which approach is more suitable when machine learning models are applied.

The motivation behind conducting a principal component analysis, and using the complete set of features, was to take a more agnostic and data-driven approach. Psychologists often carefully select meaningful variables to include in their models to get accurate predictions. The need for such selections makes it challenging to build accurate predictive models without a priori knowledge of the phenomena being studied, which is often challenging with large datasets [155]. The process can also be highly time-consuming, so investigating if a more agnostic approach can match, or exceed, the predictive ability of aggregated features is a topic of interest.

We note that by constructing the principal components, we reduce the overall explainability of the models. The fact that the components are linear combinations of the original features adds an extra layer of complexity to all the models, even though the linear combinations are interpretable. There might be a tradeoff between explainability and a reduction of the work associated with the feature selection process or predictive ability. If constructing the principal components improves prediction, a choice has to be made regarding what is most important for the researcher, to be able to explain the predictions made or to achieve high predictive abilities.

8.3.2 Procedure: Following a Supervised Learning Framework

The whole modeling process, from processing raw data to optimizing the different models, is outlined in this section. We present the process following the steps from the framework in chapter 4. Specifically, four steps were described: handling data, model assessment, model selection and model optimization.

Handling Data

Before any modeling could occur, the data had to be processed from a raw format to a complete dataset. The following subsections describe the steps taken in this process.

Step 1: Participant Selection and Available Features

A participant selection process was conducted where women with responses to all five questionnaires and who participated in the study for the first time were selected. Out of the 51 170, women that had answered all five questionnaires, 9363 were removed due to participating with several pregnancies. The first pregnancy was kept in the sample. Each data point thus belonged to a unique woman, and the total number of participants was 41 807. Figure 8.3 depicts the participant selection process, and Table 8.2 displays some characteristics of the sample from the raw data. For each dataset created, the data were split into a training and test set, with 80% of the data being used for training and the remaining 20% for testing. The split was random, but all the different training and test sets contain the same unique individuals across datasets.

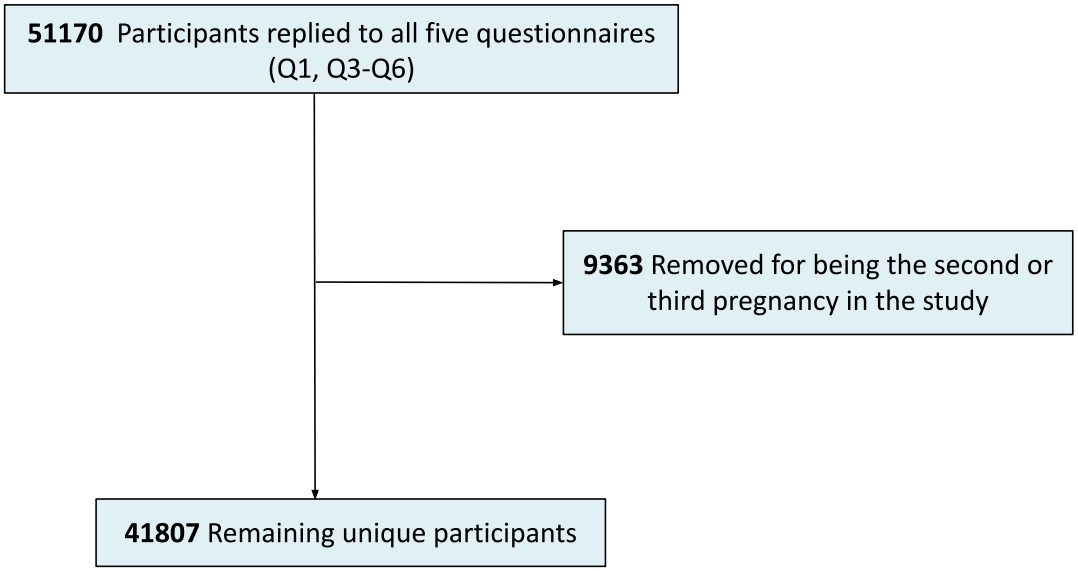

Figure 8.3: Flow-chart of the participant selection process for our methodological comparison over five time points.

The existence of different versions of the questionnaires, and the fact that the minority of the women that had partaken in the study had answered version A of the questionnaires, led us to remove all items that appear exclusively in this version. If an item had been reformulated between two versions, meaning that there exist two unique item IDs for the same question, the two were combined into one under the ID with the highest number of respondents.

As described in section 8.1, there are over 7000 single items in all of the questionnaires combined. However, only 1323 items were made available to us due to privacy regulations. The items not present in the available data were identified from the instrument documentations in [131] and are mostly related to medication, diet and previous physical health problems. A more detailed description of the missing data is given later in the chapter. The complete list of all the single items used in the different analyses is included in Appendix D.

Step 2: Transforming Data

The majority of the data in the MoBa dataset is either ordinal or ratio data. There are also some binary nominal items having the answering options of "no" and "yes", encoded as 1 and 2, respectively. No transformations were made for the ordinal and ratio data, given that the ordinal data was integer encoded, while the binary variables were shifted from (1/2) to (0/1) for convenience. All of the input data were standardized as described in subsection 4.2.2, and this step is crucial in order to investigate regression coefficients in the linear model and elastic net.

Table 8.2: Selected characteristics of the 51 170 women that had answered all five questionnaires in the MoBa study.

Characteristics	N	%
Completed Education		
Primary (9 years)	670	1.3 %
Secondary (12 years)	11824	23.1 %
Higher Ed. ≤ 4 years	17265	33.7 %
Higher Ed. > 4 years	10135	19.8 %
Missing	11276	22.0 %
Mother's Income (in NOK)		
No Income	729	1.4%
> 200 000	9809	19.2%
[200 000, 300 000)	14016	27.4%
[300 000, 400 000)	10929	21.4%
≤ 400 000	5118	10.0%
Missing	10569	20.7%
Social Support		
Yes	40262	78.6 %
No	1256	2.5 %
Missing	9652	18.9 %
Norwegians as Native Language		
Yes	36957	72.2 %
No	4166	8.1 %
Missing	10047	19.7 %
Parity		
0	15294	29.9 %
≤ 1	26044	50.9 %
Missing	9832	19.2 %
Living with Partner		
Yes	40345	78.8 %
No	10825	21.2 %

If the data is not standardized, the coefficients depend on the scale of the input data. It should be mentioned that the standardization was not applied to the data until after it was imputed.

Step 3: Outliers and Missing Values

A substantial percentage of the responses in the MoBa dataset were encoded using a Likert scale. Hence, identifying responses with values that lie outside the accepted integer interval was easy. All responses identified as laying outside this interval were encoded as missing.

A high number of items have a non-response as a valid response, leading to a lot of missing data points that were not missing. Table 8.3 illustrates how the problem manifests itself in the raw data, with the hyphen representing non-responses. The encoded responses in Table 8.3 are arbitrary and do not represent an actual participant in the study.

Table 8.3: The table illustrate how some items give rise to missing data points by having non-responses as valid answers. The selected items are from questionnaire Q4, with the item IDs DD746-DD754.

Did you take any of the following substances during the last 3 months of your pregnancy?	No	Yes, last 3 months	Yes, after birth
Hash	1	-	-
Amphetamine	-	1	-
Ecstasy	1	-	-

To remedy the high number of missing data, all features having more than 20% of their data missing were manually inspected. A new response "no need for response" was created if deemed appropriate. In total, 603 features were inspected for a high number of missing data points.

If the response exceeded a reasonable value for the ratio data, it was set to missing. What is deemed as a reasonable value depends on the item. The following example illustrates the process. In Q1, the participants are asked to write down " The usual number of paid working hours a week before you became pregnant and at present" (item ID AA1166 and AA1167). If the response is higher than the total number of hours in one week, the data point is set to missing. This process is repeated for all items containing ratio data.

All of the data collection was done by administrating self-report questionnaires by mail, as explained in section 8.1, meaning that all of the answers had to be manually encoded by humans. This can explain why a non-negligible number of items are wrongfully encoded. Some features ended up having all negative responses after the data cleaning procedure was done. The total number of items in each questionnaire is given in Table 8.4.

Table 8.4: Total number of single items found in each questionnaire.

Available Items				
Q1	Q3	Q4	Q5	Q6
254	168	184	278	286

Step 4: Imputation

The data was imputed on an item level using MICE with regression trees. We imputed the raw data before standardizing it and before creating aggregated variables. This prevents loss of information in the observed data for participants that did not respond to all of the items [106]. Given the underlying nature of regression trees, the imputed values for a feature f will lie within the interval $[f_{\min}, f_{\max}]$ for that given feature. Here $f_{\min}$ ($f_{\max}$) is the lowest (highest) observed value of the feature f. We assumed the data were missing at random (MAR), implying that all of the missing data could be explained by the other variables in the dataset. After all of the outliers and non-responses were cleaned out from the dataset, a total of 821 features were imputed using the IterativeImputer from Scikit-Learn [156]. Unfortunately, given how the answering options were designed in the MoBa study, binary items contained a high number of missing values due to non-responses. After completing step 3 in our preprocessing procedure, only continuous and category items were left to be imputed.

Step 5: Dimensional Reduction

All of the preprocessing steps described above were applied to all of the data before the construction of the different subsets began. In the fifth and final step in our data handling process, the data received different treatment with respect to how their dimensionality was

reduced. This section specifically focuses on and describes how we constructed the different datasets used to fulfill the different research aims.

For the first numerical experiment, we created three different subsets from three different dimensionality reduction procedures. As a reminder, only independent variables from the prenatal period were included in the three subsets. We started by creating a dataset with a very limited number of features through feature selection with aggregation, followed by a second subset consisting of all the single items belonging to the aggregated features in the first dataset. The third was a dataset constructed from the principal components of the complete dataset. A fourth dataset containing all available features from the prenatal period was also included in our analyses. We created all of these datasets to be able to fulfill research aim 1.4.

For the second experiment, where we were interested in predicting levels of anxiety and depression using exposures measured concurrently at 6, 18 and 36 months postpartum, three datasets were created. Each dataset consists of independent variables from one of the three postpartum time points.

Feature Selection with Aggregation

From Table 8.4, there were a total of 421 items collected during the prenatal period. We selected features from these 421 variables based on the background information regarding risk factors in chapter 2. All items assessing the selected risk factors were included in the new subset.

One of the instruments included in the first questionnaire was the lifetime history of major depression scale. This scale had eight items closely related to the DSM-III criteria for major depression. For it to be the met, there are four conditions: i) three types of symptom items are endorsed, ii) one of these has to be the symptom of "feeling depressed", iii) three types of symptoms coincide and iv) some externally negative incident did not cause the depression. Based on the four conditions from the DSM-III, the new aggregated feature was created by combining three (AA1572, AA1577 and AA1579) out of the eight items into one binary variable. The first item addresses the symptom of feeling depressed. The second asks if the participant had at least three problems simultaneously, and the third item asks if there was a particular reason for this. If the participant answered positive to the two first and negative to the third item, they met the four DSM-III criteria and were given the value 1 in the new aggregated feature. If the criteria were not met, the variable was encoded as 0.

Items measuring partner relationship satisfaction from the scale with the same name were aggregated into one variable by taking the mean of all the ten items. The same treatment was given to the satisfaction with life scale. Since it was measured in both Q1 and Q3, two aggregated variables were created from the mean of the five items.

To account for possible psychological disturbance during the pregnancy, we included items that measured social support, exposure to abuse and adverse life events and socioeconomic status through education- and income level and immigration status. The aggregated feature for social support was a binary variable created by following the same treatment as it was given in [20]. Here it was measured through one item from Q1, "did the participant have anyone other than their partner they could ask for advice in a difficult situation?" (AA1545). The item had three response options: i) no, ii) yes, 1-2 persons, and iii) yes, more than 2. If answered positive (negative) to any of the two last response options, the participant was encoded as having social support and given the value 1 (0).

We let the items from Q3 serve as a baseline for the new aggregated binary feature describing previous experiences of physical and sexual assault. There were, in total, four questions about emotional-, physical- and sexual abuse. To each question, there were three responses: i) no, never, ii) yes, as a child and iii) yes, as an adult. There were two additional questions about who the perpetrator was and if it happened during the last year for each question. If the participant answered yes to experiencing any type of abuse, regardless of experiencing it as a child or adult, the binary variable is given the value 1.

An aggregated variable was created from the 11 items in Q3 describing adverse life events. The aggregated variable was binary, and if the participant had answered yes (no) to any of the 11 questions, the variable was given the value 1 (0). The two items related to the mother's education level were used to create an aggregated variable in the same manner as in [47], where the educational status of the participant is encoded in a binary variable. If she finished or started higher education, the value of the feature is 1. In the case where the highest degree of education is high school or less, the value is 0.

Income level was not aggregated into a new feature. Instead, the two items (AA1315 and AA1316) related to the parents' yearly gross income in Q1 were kept in their original form. Immigration status was accounted for in the same way as in [47] by including an item from Q1 about the parents' native language (AA1305). If the participant or the child's father had a native language other than Norwegian, the participant was coded as an immigrant through a binary variable, having the value 1.

We included scales related to self-efficacy, negative affectivity and self-esteem in the aggregated feature dataset to capture the different personality characteristics of the participants. The aggregated variable constructed from the general self-efficacy scale was the mean value of the five instrument items. Three items from the 5-point anger subscale of the differential emotions scale and the short version of the Rosenberg self-esteem scale were included to account for the personality trait negative affectivity in the data. Data from the two scales were made into two aggregated features. Given that the differential emotion scale was only included in Q3, we used only the data from the third questionnaire when constructing the two aggregated features. Both scales used a Likert scale, so the aggregated features are the mean of all the items included. In addition to the risk factors described above, the mean SCL scores from the first and third questionnaires were included, making the total number of features for this subset 17.

Feature Selection without Aggregation

Next, we created a dataset that consisted of all the single items used in the aggregated features described in the preceding section. The total number of single items was 84.

Principal Component Analysis

The principal component analysis was conducted using the PCA class from Scikit-Learn [156], and the number of principal components used in the analyses made up 95% of the explained variance in the data. The number of principal components was 289, a reduction of 132 features from the complete set of predictors.

Subsets for Concurrent Time Points After Birth

These three subsets contained all single items from each specific time point after the participant had given birth. Meaning one subset was created for Q4, with all the items from the fourth questionnaire, and so on for Q5 and Q6. They were created in relation to the second research aim, specifically 2.2. For clarification: the three subsets above consisted of independent variables collected during the prenatal period, while these three subsets strictly have independent variables from the time after the child was born.

8.3.3 Model Assessment

The models' predictive ability will be assessed on the test set using the root mean squared error (RMSE) metric. This is sometimes referred to as leave-one-out cross-validation. Given the normality assumption about the residuals in a linear model, RMSE is the most suitable choice instead of other metrics such as the mean absolute value (MAE). Given the higher weighting of larger errors, the RMSE will punish high errors more than MAE. We argue that this is a

desirable trait when machine learning applying in a clinical setting. Higher prediction errors could potentially have more severe consequences than a model showing minor discrepancies across the whole domain.

An article from 2005 argued that the RMSE is an unambiguous and inappropriate metric [157]. These claims were refuted in an article by Chai and Draxler in 2014 [109]. The 2005 article by Willmott and Matsuura used four hypothetical sets of four errors to show that the RMSE was unstable in cases where the MAE was constant. However, the small sample size is not necessarily representable for real-world problems, with sample sizes often far greater than four. With this in mind, we chose the RMSE to be our metric of comparison when comparing the different models.

Utilizing this metric ensured that the error had the same physical dimension as the dependent variable. Applying this metric to all of the predictions across the numerical experiments allowed us to make objective comparisons across the different datasets. Other metrics that incorporate the degrees of freedom in the models would not be a suitable choice, given that we are comparing models with a varying number of parameters.

In accordance with research aim 1.5, "Evaluate the performance of different methods for identifying individuals at risk for clinical levels of depression and anxiety", we also assessed how the different models predicted the underlying distribution of the true targets and their residuals. This was done for the predictions made in the first numerical experiment, specifically for the predictions made when the models were trained on the principal component analysis and all available items. A low prediction error does not guarantee a good fit of the predictions to the actual values, and a visual inspection can reveal possible deviations from the true distribution. We also assessed an approximation of the model's sensitivity and specificity by conducting a post-analysis dichotomization of the predictions, using the cut-off score from the SCL-10 as a divider for our continuous target variable. We note that it is only an approximation due to the fact that the constructed binary predictions do not stem from a classification model. Had the sensitivity and specificity been calculated from predictions made by a classification model, we would expect a different result.

8.3.4 Model Selection, Optimization and Implementation

All hyperparameters belonging to the different models were determined through a 5-fold cross-validation using a random search with either ten, 15 or 30 iterations, depending on the computational cost related to the search. A unique set of hyperparameters was identified for the machine learning models for each dataset. The numerical experiments related to the first research question required 12 separate sets of hyperparameters, one set for each machine learning algorithm for each subset of data. Investigating the time dependence in the predictors, and thus partially answering the second research question, called for one set of values for each model per time point, making it a total of nine sets of hyperparameters.

We implemented the random search with the 5-fold cross-validation using the RandomizedSearchCV class from Scikit-Learn. The class instance can either take a distribution or a list of possible values for each hyperparameter. The specific details for finding the optimal hyperparameter values and how the models were optimized and implemented for each machine learning model are outlined below. All of the specific sets of hyperparameters for each numerical experiment are given in Appendix E.

Elastic Net

The elastic net was implemented using the ElasticNet class from Scikit-Learn, which minimizes the objective function

$$\mathcal{C}(\zeta, \gamma, \boldsymbol{\beta}) = \frac{1}{2N}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \frac{1}{2}\zeta(1-\gamma)\|\boldsymbol{\beta}\|_2^2 + \zeta\gamma\|\boldsymbol{\beta}\|_1, \tag{8.3}$$

where

$$\zeta = \lambda_1 + \lambda_2 \qquad \text{and} \qquad \gamma = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Minimizing this cost function is equivalent to the solving constrained problem given in section 5.2,

$$\hat{\boldsymbol{\beta}}_{\text{EN}}(t) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}:(1-\alpha)\|\beta\|_1+\alpha\|\beta\|_2^2\leq t}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2, \qquad \text{(5.30 revisited)}$$

where $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$. So instead of searching for optimal values for λ_1 and λ_2 explicitly, the search was conducted with respect to ζ and γ. Be aware of that in the code documentation for the ElasticNet class, the two hyperparameters are named differently, with ζ being `alpha` and γ is `l1_ratio`. When conducting the randomized grid search, the possible values for the two hyperparameters ζ og γ were given as two uniform distributions between $[10^{-4}, 1]$ and $[0, 1]$, respectively. Given the low computational cost of conducting a grid search for the elastic net, 30 iterations were performed.

We used the default number of 1000 iterations to fit the regression coefficients during the optimization process. The ElasticNet class fits the intercept by default, and the tolerance ϵ has the default value 10^{-4}.

Neural Network

We implemented a dense neural network using the Keras API for Tensorflow 2 [158]. When tuning hyperparameters, a Scikit-Learn wrapper was utilized to take advantage of the RandomizedSearchCV class. Different values for the number of hidden layers, nodes in each layer, batch sizes, epochs, activation functions, optimizers, learning rate and different regularizers were explored. Due to high time consumption, only ten random samples were performed during the search. The loss function minimized during both the tuning and training process was the RMSE.

A neural network can take an arbitrary number of nodes and hidden layers, leaving you with an infinite number of possible combinations when designing a network. We restricted our search for the optimal network architecture to a maximum of six hidden layers, with combinations of $10, 100, 500$ and 1000 nodes in each layer. The learning rate could take the value 10^{-3} or 10^{-2}, while the batch size was either 32 or 64. We tried the ReLU and hyperbolic tangent for the activation function, and the number of epochs could be 100, 500 or 1000. For the regularization, we tried with- and without weight decay (or L_2 regularization) and dropout. The dropout rate was set to be either 0.2 or 0.5. The regularization parameter, λ, in the weight decay layer from the Keras API has the default value of 0.01. We did not experiment with different parameter values, and the default value was used. If dropout was included as an active regularizer, a dropout layer was added between every other hidden layer. Two optimizers were included in the search, stochastic gradient descent and ADAM. If the ADAM optimizer was chosen, the default values for its hyperparameters were applied.

During the optimization, we used the ModelCheckpoint callback from Keras to save the weights belonging to the model that yielded the lowest prediction error on the validation set as our early stopping routine. Despite being more computationally expensive, we circumvent the possibility of being stuck in a local minimum early on in the training procedure. The model's performance was evaluated on a validation set during training.

Extreme Gradient Boosted Regression Trees

The extreme gradient boosted regression trees, also called the XGBoost model, were implemented through the XGBoost library [127]. A Scikit-Learn wrapper was applied to conduct the randomized grid search, and we explored different combinations of values for the hyperparameters `max_depth`, `min_child_weight`, `subsample`, `colsample_bytree`, number of

estimators and learning rate. The optimal values of hyperparameters were determined by minimizing the RMSE.

The `max_depth` parameter, describing the maximum depth of a tree, could be between two and seven nodes deep. The number of samples present in one node in order to stop partitioning the data, the `min_child_weight`, could be in the range $[1, 5]$. The learning rate was either 0.01 or 0.005. We explored the values $\{i/10\}$ for $i \in [1, 5]$ for the two parameters describing the sample and feature subsampling ratio, `subsample` and `colsample_bytree`. The number of estimators, or trees added to the ensemble, could either be $1000, 5000, 10\,000$ or $20\,000$.

The number of early stopping rounds during the training procedure was one-quarter of the total number of boosting rounds. An evaluation set was used to monitor the model's loss during training during the optimization process.

8.4 Code Availability

All of the code used in this project is made publicly available on GitHub (see https://github.com/marialinea/predicting-depression-and-anxiety-moba). The repository contains all of the scripts used to handle the data, make predictions and process the results.

Part III

Results & Discussion

Chapter 9

Results and Analysis

Here we present our findings from the experiments detailed in subsection 8.3.1. All tables showing prediction errors made by the multiple linear regression models, elastic nets, neural networks and XGBoost models contain bolded entries representing the lowest errors obtained in each specific scenario. The error is quantified through the root mean squared error. Predictions were always made on out-of-sample data if not specified otherwise.

9.1 Experiment 1: Investigating Predictive Ability and Feature Importance using Prenatal Exposures

Here we present the results associated with the first numerical experiment. With four different datasets, we predicted the mean SCL score at three time points, Q4, Q5 and Q6, which corresponds to 6, 18 and 36 months after giving birth. Depending on which version of the SCL instrument answered, the cut-off score that indicated mental illness varies. To give an indication of the number of participants that are depressed, we used the most conventional cut-off score; a mean value above 1.75 indicates a mental disorder in need of treatment. In the training sets, the prevalence of scores above 1.75 for the different three time points were $14.8\%, 15.5\%$ and 18.5%, compared to $7.8\%, 9.5\%$ and 10.8% in the test sets, for Q4, Q5 and Q6 respectively.

9.1.1 Predictions Errors

We present the prediction errors made by the different supervised learning algorithms on all four datasets, starting with the dataset containing the least amount of features. The results are directly related to the research aims

1.1 Predict levels of anxiety and depression at 6, 18 and 36 months postpartum using prenatal exposures measured at 17 and 30 weeks of pregnancy.

1.3 Compare the performance of models using i) aggregate scores on established scales, ii) item-level analyses, iii) dimensional reduction by principal component analysis, and iv) data without dimensional reduction.

1.4 Evaluate the performance of different methods for identifying individuals at risk for clinical levels of depression and anxiety.

Table 9.1 displays the RMSE when selected items were aggregated into 17 new features.

Table 9.1: Results from the 17 aggregated features from Q1 and Q3 to predict the mean SCL score six, 18 and 36 months after birth, corresponding with Q4, Q5 and Q6 respectively.

Method	Q4	Q5	Q6	Mean
Multiple Linear Regression	0.3982	0.3927	0.4434	0.4114
Elastic Net	0.3975	0.3922	0.4428	0.4108
Neural Network	0.3760	0.3828	0.4216	0.3934
XGBoost	**0.3595**	**0.3751**	**0.4135**	**0.3827**

The XGBoost exhibited the lowest prediction errors when the features were aggregated. The relative improvement to the linear regression model was 7.0%. The mean performance of the mean RMSE for all models was 0.3996.

The prediction errors from when the single items constituting the aggregated features were used as training data are shown in Table 9.2.

Table 9.2: Results from using the 84 single items that made up the aggregated features from Q1 and Q3 to predict the mean SCL score six, 18 and 36 months after birth, corresponding with Q4, Q5 and Q6 respectively.

Method	Q4	Q5	Q6	Mean
Multiple Linear Regression	0.3903	0.3897	0.4399	0.4076
Elastic Net	0.3827	0.3818	0.4301	0.3982
Neural Network	0.3745	0.3804	0.4181	0.3910
XGBoost	**0.3577**	**0.3698**	**0.4039**	**0.3771**

The lowest prediction errors belonged to the XGBoost model with an overall mean RMSE value of 0.3771. Compared to the linear regression model, this is a relative improvement of 7.8%. All models showed an improvement in their mean prediction error over all time points compared to Table 9.1, and the total mean performance of all models was 0.3934.

The third dataset was constructed from the principal components that explained 95% of the variance in the training data, and the prediction errors made by the four models are shown in Table 9.3. The total number of principal components was 289.

Table 9.3: Root mean square errors of predicted mean SCL score for Q4, Q5 and Q6, when the models were trained on the 289 principal components. Together they explained 95% of the variance in the data.

Method	Q4	Q5	Q6	Mean
Multiple Linear Regression	0.3775	0.3885	0.4267	0.3975
Elastic Net	0.3700	0.3822	0.4171	0.3897
Neural Network	**0.3558**	**0.3708**	**0.4011**	**0.3759**
XGBoost	0.3641	0.3795	0.4134	0.3856

When the principal components made up the training data, the neural network outperformed the other models and had an overall mean RMSE prediction error of 0.3729. This was 6.2% better than the linear model. An overall improvement in the mean prediction error was observed in three out of the four models, with the XGBoost model experiencing a 2.3% decline in RMSE from when the data consisted of the single items from the aggregated features. The overall performance of all the models on this dataset, quantified through the mean RMSE on all time points, was 0.3864.

The final dataset included all available items from Q1 and Q3, and the prediction errors made when predicting the mean SCL at Q4, Q5 and Q6 are displayed in Table 9.4. The total number of features was 421.

Table 9.4: Results from using all 421 predictors from Q1 and Q3 to predict the mean SCL score six, 18 and 36 months after birth, corresponding with Q4, Q5 and Q6 respectively.

Method	Q4	Q5	Q6	Mean
Multiple Linear Regression	0.3923	0.3927	0.4412	0.4087
Elastic Net	0.3674	0.3804	0.4140	0.3872
Neural Network	**0.3545**	**0.3694**	**0.4030**	**0.3756**
XGBoost	0.3598	0.3724	0.4073	0.3798

It is evident from Table 9.4 that the neural network exhibited the lowest prediction error on average. They were 8.1% lower than the errors made by the multiple linear regression models. Compared to the predictions made in Table 9.3 only the XGBoost- and the multiple linear regression model experienced an improvement in prediction errors. The mean RMSE across all time points for all models was 0.3878.

We visualized how the mean RMSE over all time points changed as a function of the increasing number of features in the training- and test data for all four models are visualized in Figure 9.1.

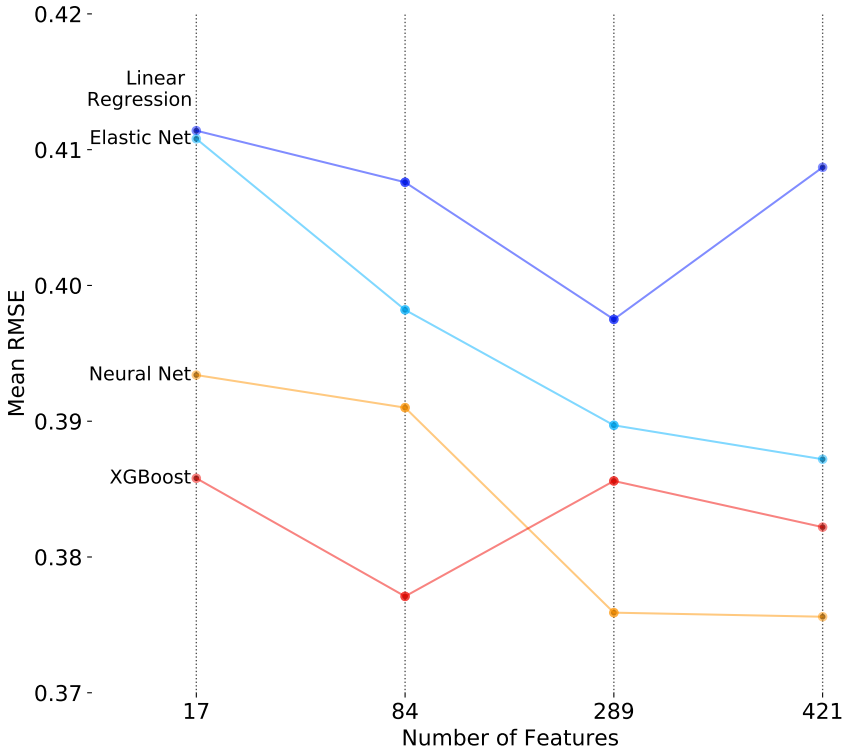

Figure 9.1: The figure displays how the mean RMSE over all time points changed when the number of features included in the training- and test data increased. When the number of dependent variables was 17, the features were aggregated, and the 289 features correspond with the principal components making up 95% of the explained variance.

Looking at the combined mean performance of all models in each dataset, the overall lowest prediction errors were obtained when the data consisted of the principal components, with the dataset containing all available items as runner up. This was also reflected in the overall mean RMSE of all the models across time points, which were 0.3864 and 0.3878, respectively. When it came to how effectively the models could predict the mean SCL score as time passed since birth, they all showed the same pattern: prediction errors increased with time.

9.1.2 Identifying Clinical Levels of Depression and Anxiety

In efforts to evaluate how the different methods were able to identify individuals at risk for clinical levels of depression and anxiety (ref. research aim 1.5), we investigated the distribution of predictions and how they corresponded with the true underlying distribution. The residuals were also graphically displayed, and we approximated the models' sensitivity and specificity. This was done on the two datasets that yielded the lowest prediction errors in the preceding section, the principal component subset and the complete dataset.

The distribution of the predictions and true values are shown in Figure 9.2.

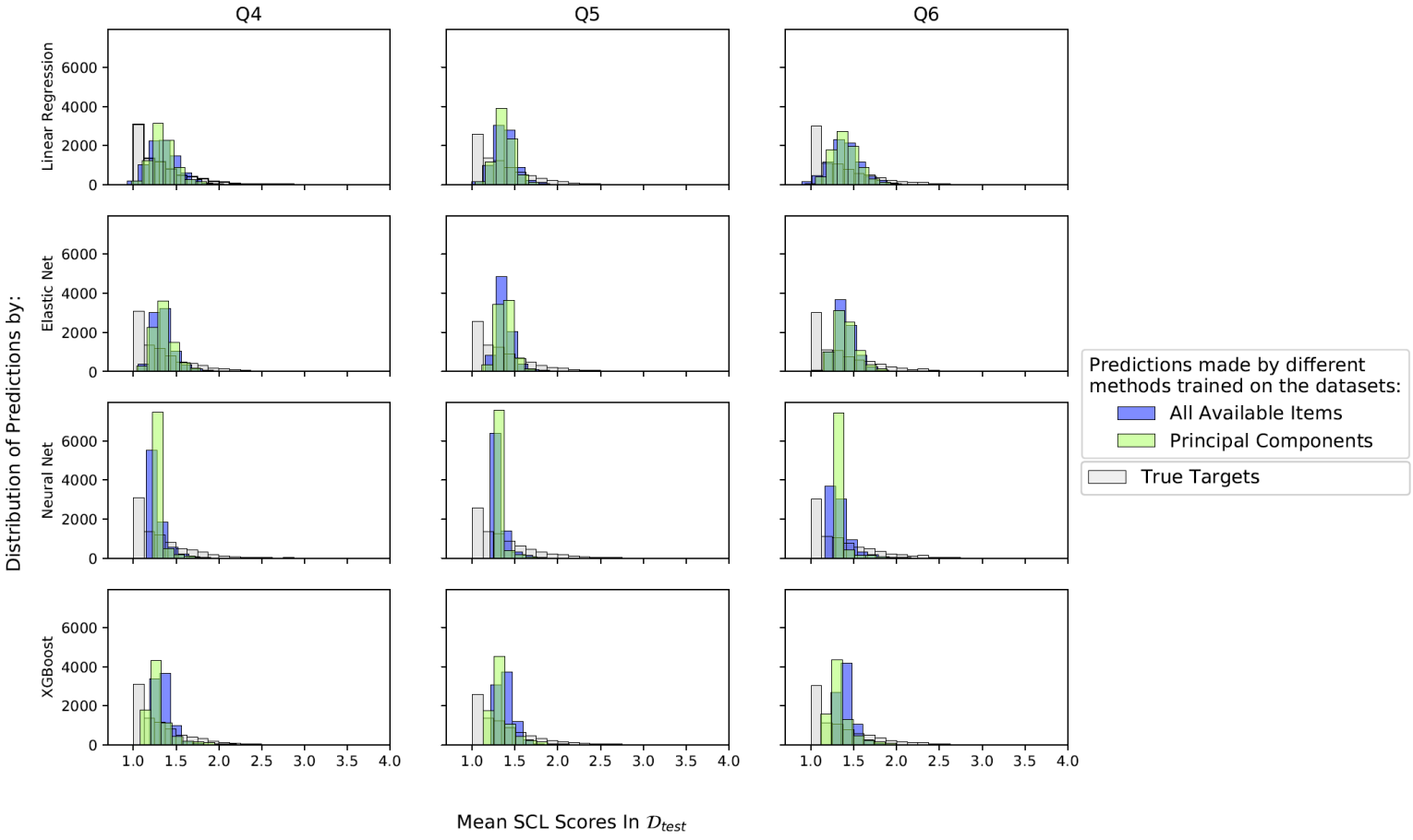

Figure 9.2: Distribution of the predictions made by the models trained on the complete set of 421 features available from Q1 and Q3 and the principal components, compared to the true targets in the test set, $\mathcal{D}_{\text{test}}$.

From Table 9.3 and Table 9.2 it is evident that the neural network exhibits the lowest predictions errors. By inspecting Figure 9.2, the neural network obtains such results by almost exclusively predicting the one value centered around ~ 1.25. The mean value of the dependent variable at all three time points in the training- and test sets was (train, test, Q): (1.35, 1.28, Q4), (1.38, 1.31, Q5) and (1.40, 1.32, Q6). So the neural networks are simply predicting the mean value

of $\mathcal{D}_{\text{test}}$. The remaining three models all exhibited some problems with predicting the tail of the true distribution, but not to the same extent as the neural network. The highest frequency of predictions falls close to the mean value for all models, i. e. they all show the highest peak around the mean of the target variable in the test set.

The residuals were plotted in order to identify which prediction belongs to which target. This not possible from the distributions in Figure 9.2. Figure 9.3 graphically displays the empirical cumulative distribution of the relative residuals. It is evident from Figure 9.3 that the multiple

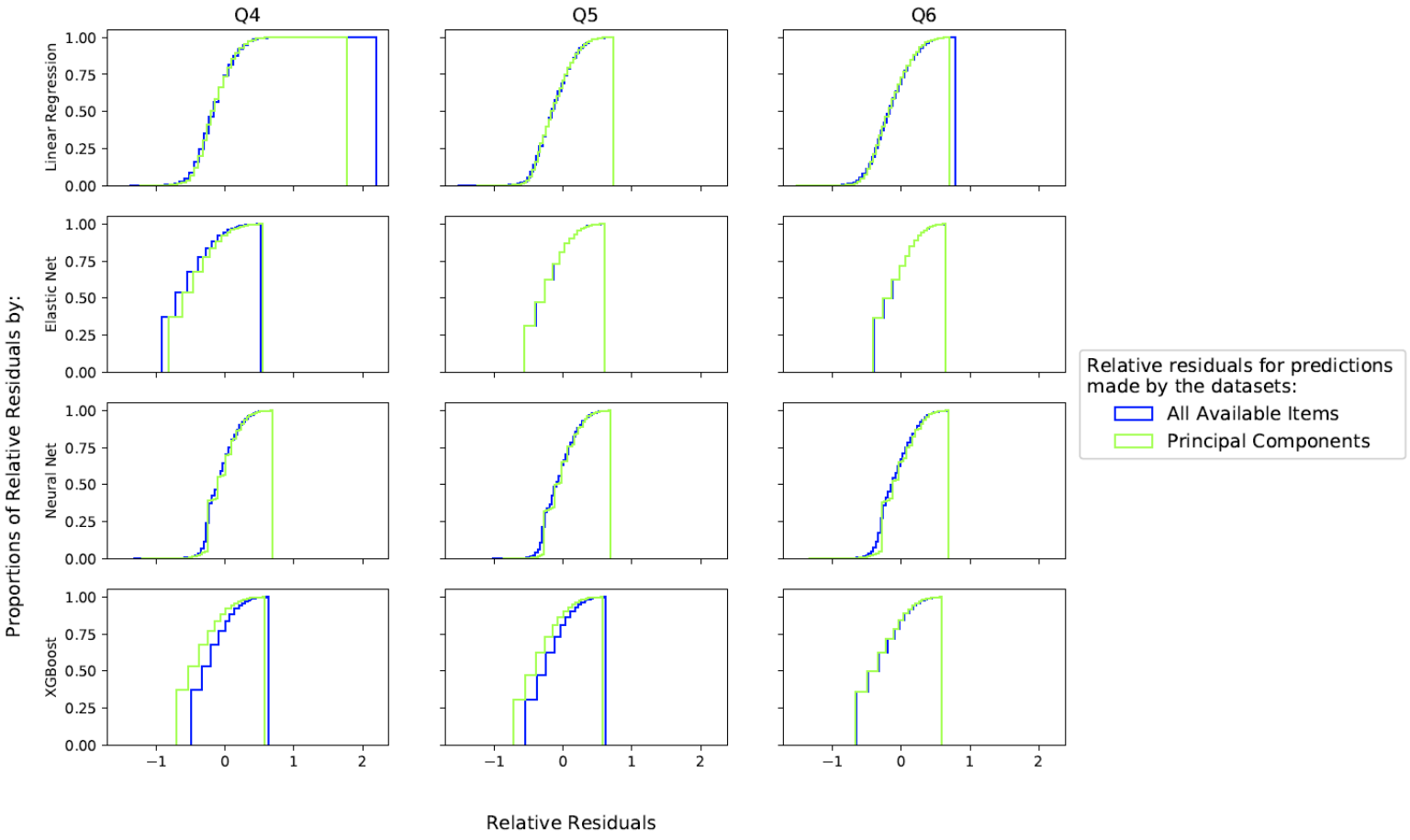

Figure 9.3: Empirical cumulative distribution of the relative residuals made on the test set, from the models trained on the complete set of 421 features available from Q1 and Q3 and the principal components

linear regression model and the neural network have somewhat normally distributed residuals. On average, each of their predictions was closer to hitting the targets compared to the elastic nets and the XGBoost models.

To approximate the models' sensitivity and specificity, we categorized the predictions made by the models into two groups, non-clinical- and clinical levels of depression and anxiety. The cut-off score for the SCL-10 (1.85) was used to separate the predictions. The mean approximated sensitivity and specificity across time points for the two datasets are displayed in Table 9.5. The low sensitivity in Table 9.5 indicates that the models failed to identify individuals showing clinical levels of depression and anxiety and incorrectly predicted a low mean SCL value when the true value was above the threshold. The multiple linear regression models both have over 55% probability of correctly identifying clinical levels when the true mean SCL score is above 1.85. The high specificity in all models describes low chances of falsely predicting mean SCL scores above the cut-off score, so-called false positives. When the XGBoost model was trained on the principal components, the sensitivity experienced an 11% relative increase.

Table 9.5: The mean sensitivity and specificity across the three time points are given for when the models are trained on all available items and the principal components. The linear models clearly exhibits the highest sensitivity in both cases, while the XGBoost models experience a relatively high increase in sensitivity when the models are trained on the principal components.

Method	All Items		Principal Components	
	Sens. (%)	Spec. (%)	Sens. (%)	Spec. (%)
Multiple Linear Regression	56.0	98.6	55.6	98.4
Elastic Net	3.9	99.4	4.3	99.4
Neural Network	3.4	99.6	2.3	99.7
XGBoost	0.8	99.9	9.6	98.0

9.1.3 Further Exploration of the Predictions

We had not expected the low sensitivity measures in Table 9.5 or the behaviour of the prediction distributions in Figure 9.2. In an attempt to better understand the underlying reasons for these behaviours, the absolute correlation of the independent variables with the target variables was inspected. A possible explanation, or hypothesis, could have been that the original dataset contained many insignificant features that could have interfered with the training procedure and inhibited learning. The empirical cumulative distributions of the absolute correlation among the 421 features with the dependent variables for all three time points are shown in Figure 9.4.

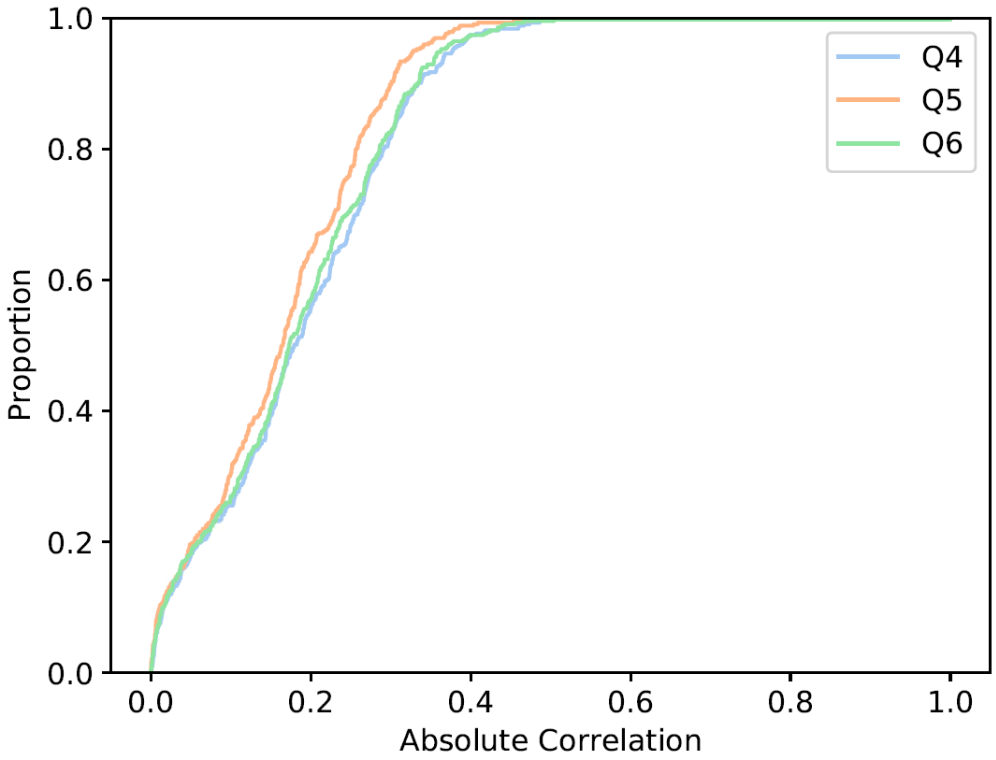

Figure 9.4: Empirical cumulative distribution of the features' absolute correlation with the dependent variables at the different time points. From the plot it is evident that the maximum absolute correlation is ~ 0.5 for a small subset of features in all time points.

In Figure 9.4, all three target variables show roughly the same amount of (absolute) correlation with the independent variables. Half of the features have an absolute correlation equal to or less than 0.2, and the highest correlation is around 0.5. Based on the relationships between

the dependent- and independent variables, it appears plausible that there is an underlying structure for the models to learn to make meaningful predictions. However, the number of features that exhibits a low correlation with the targets may still affect the training procedure.

We further investigated our hypothesis by creating a fifth dataset, strictly containing features with an absolute correlation of 0.35 or higher with the target variable. The number of features for each time point varied, and Q4 had 33 "high" correlating features, Q5 had 14 and Q6 with 28. The number of new subsets created was thus three. However, they were all constructed for the same purpose: investigating if the number of low correlated features interfered with the models' learning ability. The predictions errors made by the four models are shown in Table 9.6.

Table 9.6: Results from using only features that had an absolute correlation of 0.35 or higher with the target variable at the different time points. The table shows the prediction error made when the number of features was 33, 14 and 28 for predicting the mean SCL at Q4, Q5 and Q6 respectively.

Method	Q4	Q5	Q6	Mean
Multiple Linear Regression	0.3645	0.3802	0.4412	0.3953
Elastic Net	0.3612	0.3793	0.4201	0.3868
Neural Network	0.3700	0.3905	0.4174	0.3926
XGBoost	**0.3481**	**0.3671**	**0.3998**	**0.3716**

The prediction errors made by the XGBoost model outperformed all other prediction errors in the four tables in subsection 9.1.1. Compared to the linear model, the relative improvement in prediction error for the XGBoost model was 6.0%.

As previous results have shown, low prediction errors do not necessarily equal a desirable outcome. The relative residuals of the predictions made on the correlated features was plotted, and added to Figure 9.3, creating Figure 9.5. From Figure 9.5, it is evident that using highly correlated features did not consistently produce lower residuals or more normally distributed residuals. This indicates that the number of features and their correlation with the target variable can not fully explain the behaviour observed in Figure 9.2. This led us to question the models: are they able to identify patterns in the training data but struggling to generalize to the test set, or are they having problems learning in general? In order to answer this, the trained models were evaluated on $\mathcal{D}_{\text{train}}$. The prediction errors for the principal components dataset are shown in Table 9.7, and the errors for the complete dataset in Table 9.8.

Table 9.7: Prediction errors, quantified with the RMSE, made on the training set by the models when they were trained on the principal components. The errors from the test set are repeated for comparison, and can be found in Table 9.3.

Method	Q4		Q5		Q6		Mean	
	Train	Test	Train	Test	Train	Test	Train	Test
Multiple Linear Regression	0.3337	0.3775	0.3497	0.3885	0.3781	0.4267	0.3583	0.3957
Elastic Net	0.3352	0.3700	0.3510	0.3822	0.3794	0.4171	0.3552	0.3897
Neural Network	0.3571	0.3522	0.3793	0.3655	0.4054	0.4009	0.3806	0.3729
XGBoost	0.2363	0.3641	0.2462	0.3795	0.2712	0.4134	0.2512	0.3856

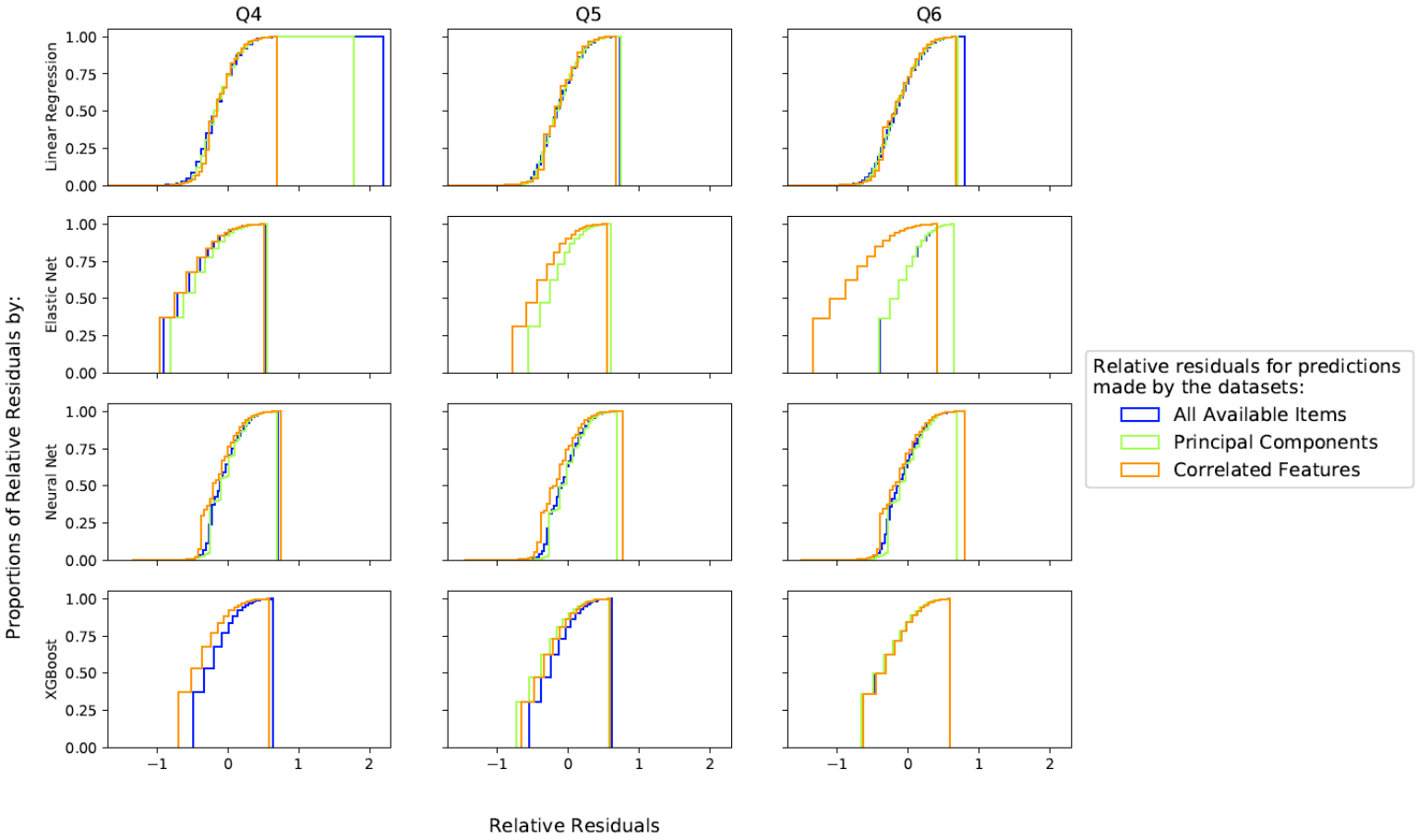

Figure 9.5: Empirical cumulative distribution of the relative residuals made on the test set, from the models trained on the complete set of 421 features available from Q1 and Q3, the principal components and the features having a correlation higher than 0.35 with target variables.

Table 9.8: Prediction errors made on the training set by the models when they were trained on all available features. The errors from the test set are repeated for comparison, and can be found in Table 9.4.

Method	Q4		Q5		Q6		Mean	
	Train	Test	Train	Test	Train	Test	Train	Test
Multiple Linear Regression	0.3306	0.3923	0.3477	0.3927	0.3748	0.4412	0.3510	0.4087
Elastic Net	0.3367	0.3674	0.3526	0.3804	0.3814	0.4140	0.3569	0.3872
Neural Network	0.3101	0.3545	0.3383	0.3694	0.3577	0.4030	0.3353	0.3756
XGBoost	0.2866	0.3598	0.3064	0.3724	0.3225	0.4073	0.3051	0.3798

All of the models showed lower prediction errors on the training set, both Table 9.7 and 9.8, except for the neural network in Table 9.7. The empirical cumulative distribution of the relative residuals is plotted in Figure 9.6. The distributions of residuals in Figure 9.6 all exhibit a curve similar to a normal distribution, and there are no significant differences in residuals for the two different datasets. This could indicate that the models are having problems with generalization.

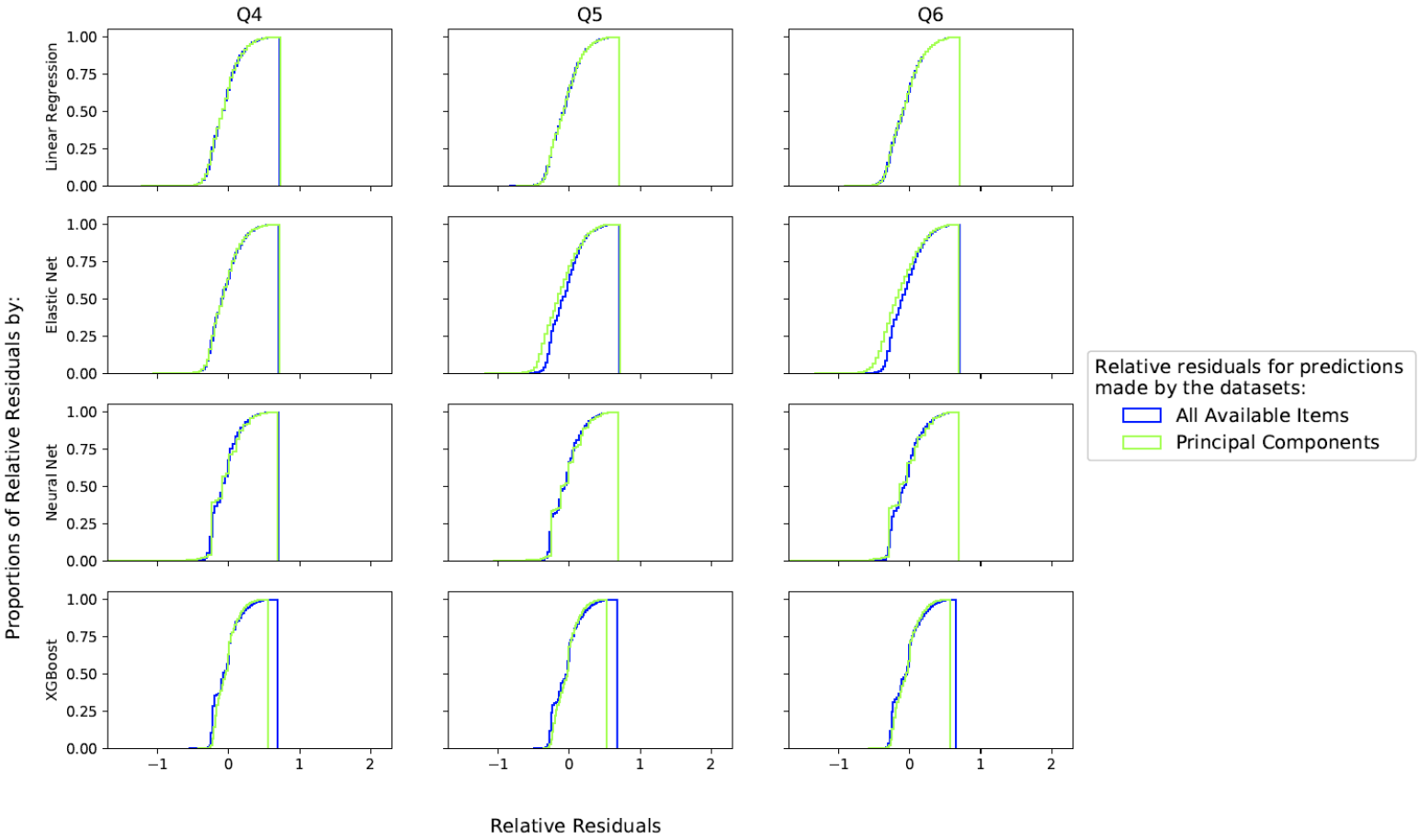

Figure 9.6: Empirical cumulative distribution of the relative residuals when the models where evaluated on $\mathcal{D}_{\text{train}}$. The figure displays the residuals for when the models where trained on the principal components and all available features.

9.1.4 Identifying Prenatal Risk Factors

To identify prenatal exposures associated with increased risk of symptoms of anxiety or depression at 6, 18 and 36 months after birth (ref. research aim 2.1), we used the predictions made when the principal components and all 421 features were used as training data to investigate several feature importance measures. Specifically, the regression coefficients from the multiple linear regression models and the elastic nets were investigated together with the gain scores from the XGBoost models. By comparing the different feature importance measures, we can also address research aim 2.3, stating " Compare the sets of variables identified through traditional linear model methods and machine learning approaches.".

The regression coefficients from the multiple linear model trained on all available features are depicted in Figure 9.7. These coefficients belonged to the linear regression model that produced the prediction errors in Table 9.4.

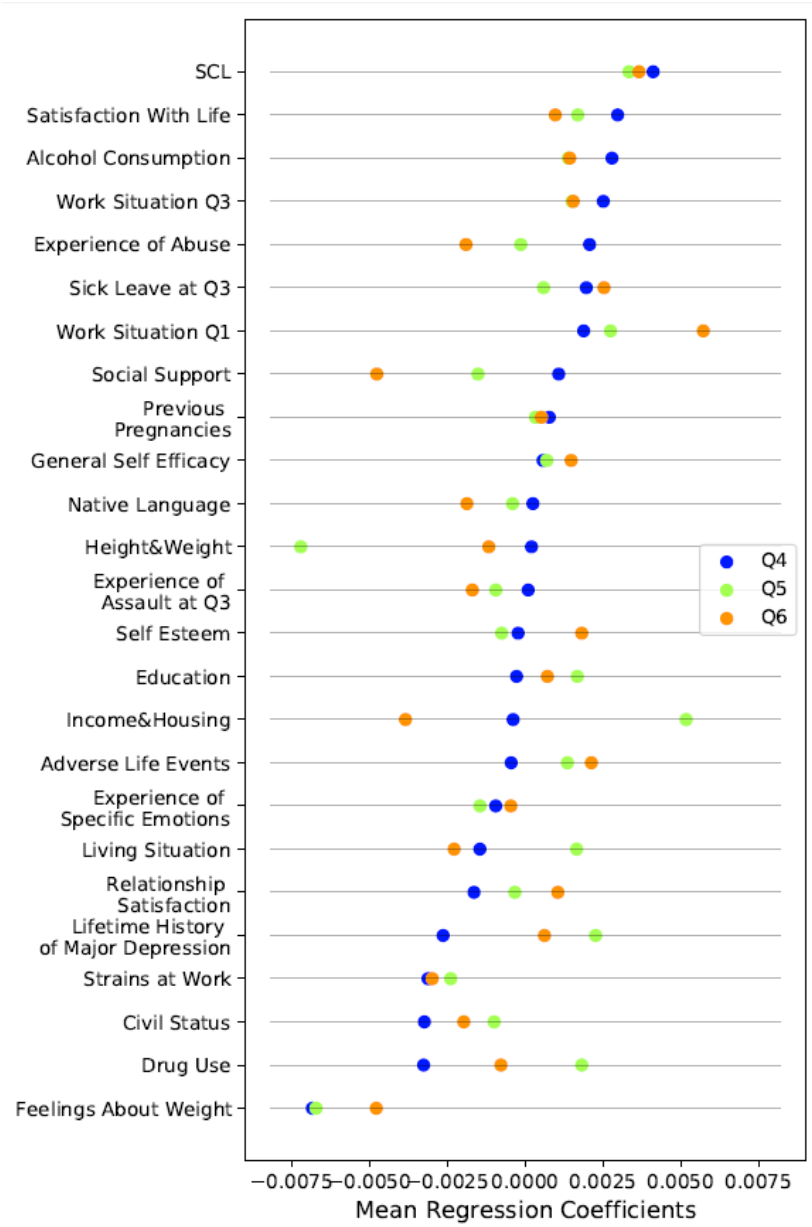

Figure 9.7: Mean regression coefficients for groups of features in the multiple linear models trained on all available features. The group named "SCL" refers to the participants' answers to the SCL instrument at Q1 and Q3.

It is evident from Figure 9.7 that some groups of features experienced a change in both sign and magnitude as the time went by, such as the features related to the experience of adverse life events, relationship satisfaction and experience of abuse. The items related to the SCL checklist in Q1 and Q3 have, not surprisingly, a consistent effect on the mean SCL score at later time points. The negative coefficients for the items related to feelings about weight indicate that a relaxed attitude towards their own weight and weight gain has a preventing effect.

The regression coefficients for the elastic net trained on the same data are shown in Figure 9.8.

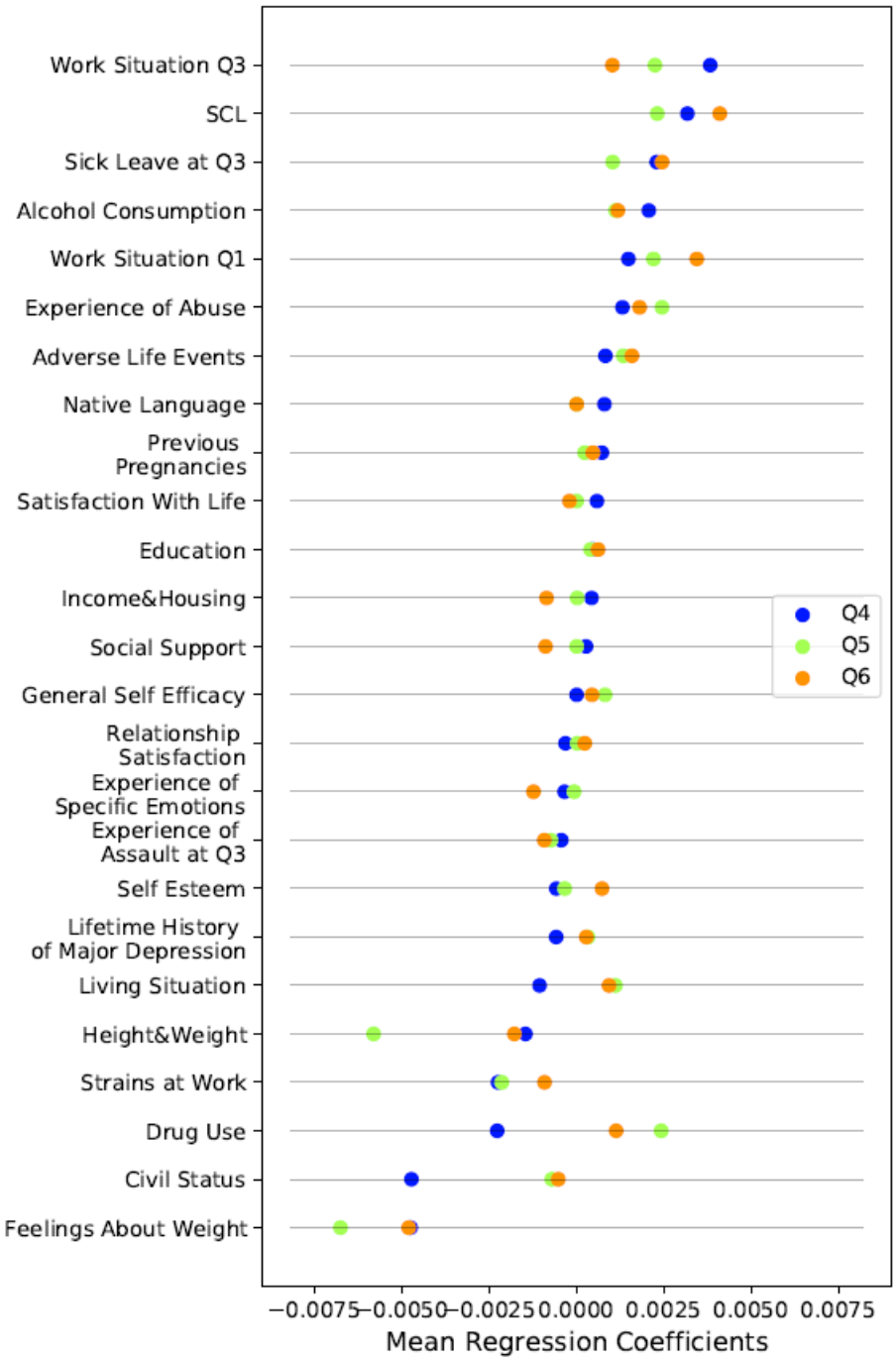

Figure 9.8: Mean regression coefficients for groups of features in the elastic nets trained on all available features. The group named "SCL" refers to the participants' answers to the SCL instrument at Q1 and Q3.

The mean regression coefficients in Figure 9.8 showed less variation over time compared to the ones in Figure 9.7. However, the group of features with the highest and lowest values are somewhat consistent between the two figures. Given the L_1-regularization term in the elastic net, several groups of items had mean values closer to zero than in the linear regression model. Figure 9.8 we observe that the coefficients related to drug use and the experience of abuse gradually increase, while the mother's work situation in Q3 becomes more insignificant with time.

The gain scores from the XGBoost model trained on all available features are given in

Figure 9.9.

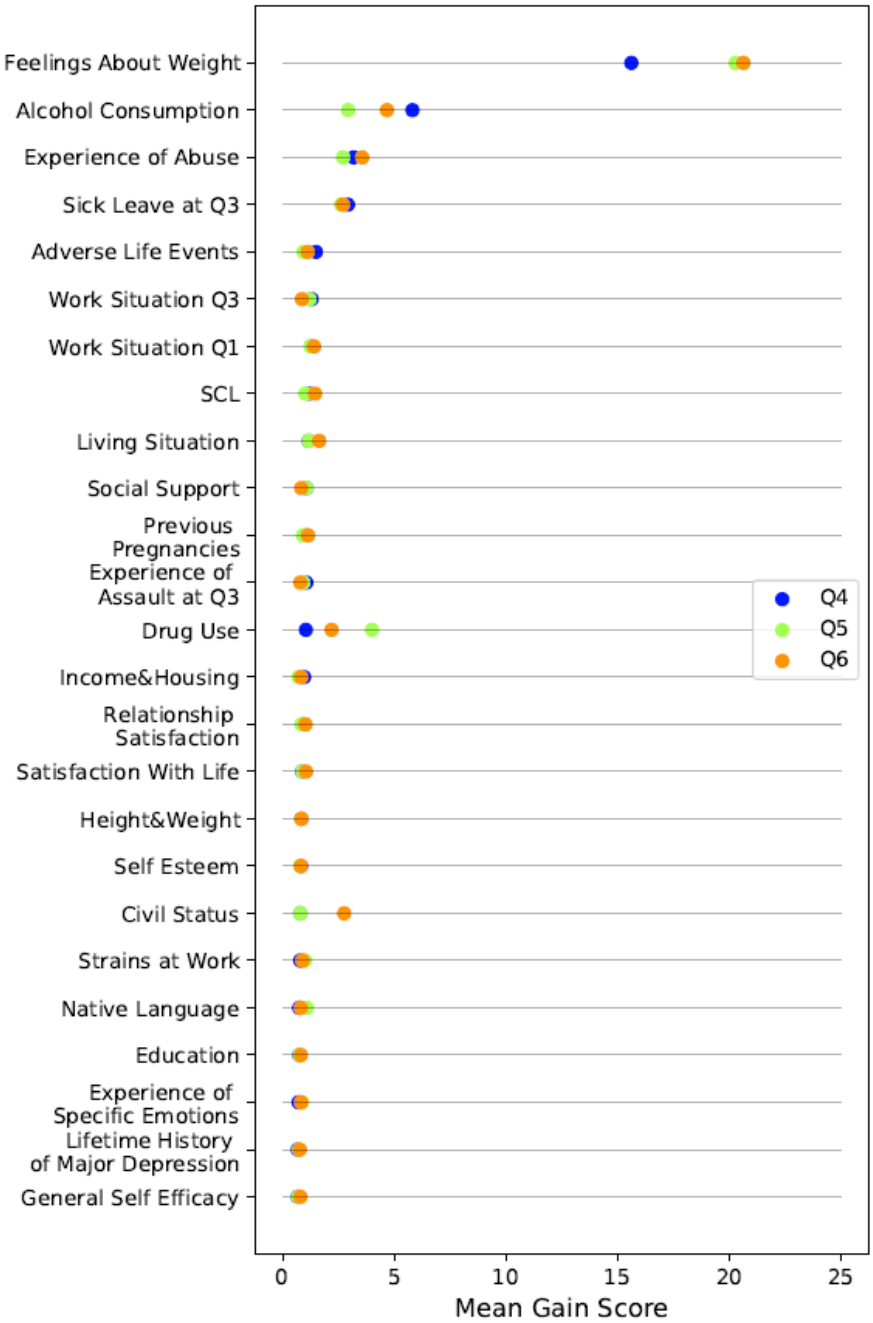

Figure 9.9: Mean gain score for groups of features in the XGBoost models trained on all available features. The group named "SCL" refers to the participants' answers to the SCL instrument at Q1 and Q3.

From Figure 9.9, it becomes clear that the items related to weight had the most considerable contributions to reducing the prediction error. There are some minor variations in which of the groups of items the three different models assigned the highest importance, but overall there seems to be a consensus between the models and how they form their predictions when trained on all available items.

As described in subsection 8.3.1, to investigate the feature importance from when the models were trained on the principal components, the 20 highest weights in the components with the highest coefficients or gain scores were grouped. The count of how often a group appeared was

used as an importance measure. The feature importance plot for the multiple linear regression model is shown in Figure 9.10.

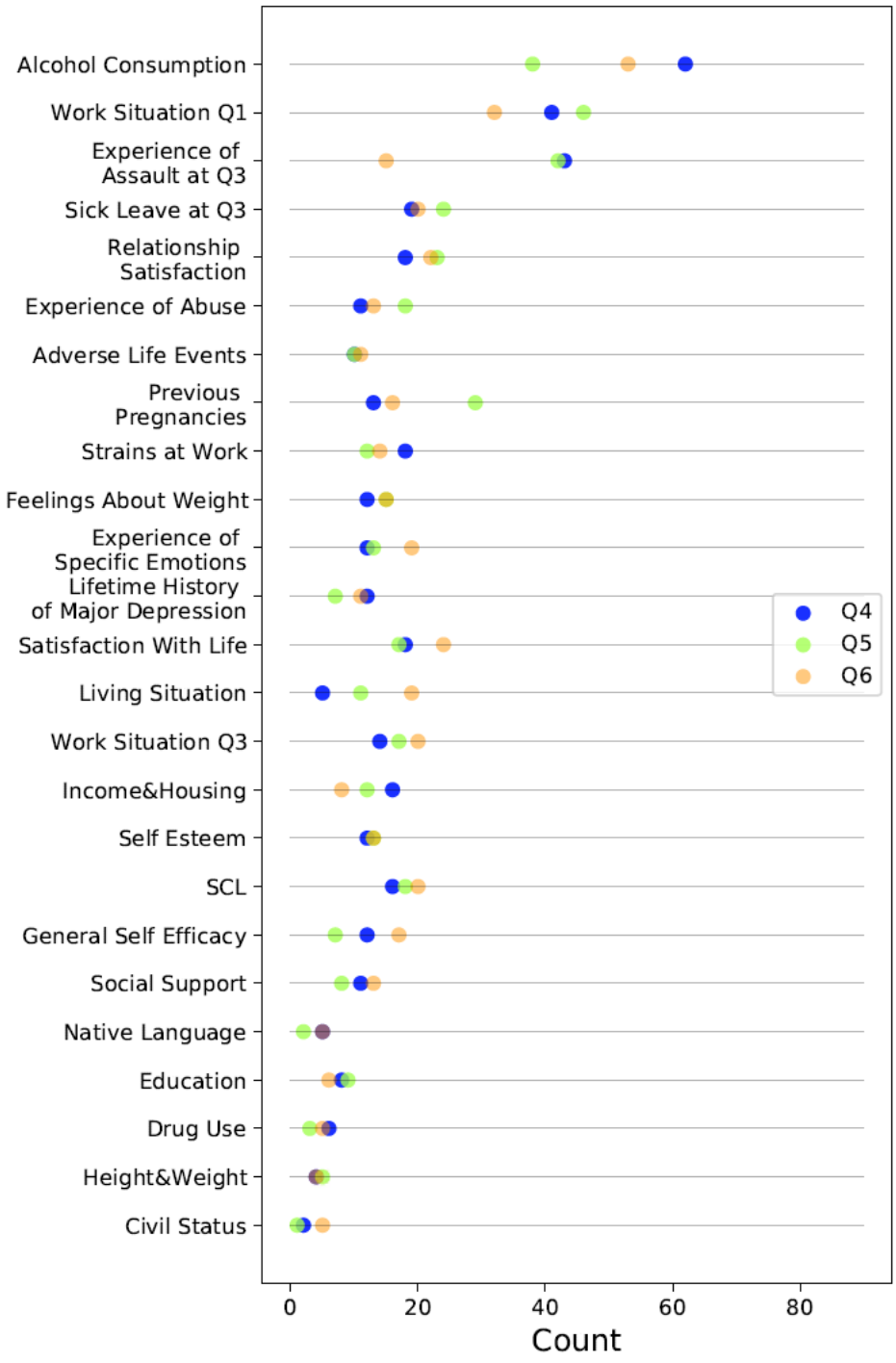

Figure 9.10: The figure displays the top 20 weights in each of the 20 principal components with the largest regression coefficients from a multiple linear regression model that predicted the mean SCL score at three time points. The prediction errors of the linear model can be found in Table 9.3. The group named "SCL" refers to the participants' answers to the SCL instrument at Q1 and Q3.

From Figure 9.10, it is evident that the single items describing behaviour related to alcohol consumption frequently appeared in the top 20 principal components. The items related to the

experience of assault at Q3 see a decline in frequency when predicting the mean SCL score 36 months after birth.

How the different groups appeared in the principal components from the elastic net is shown in Figure 9.11.

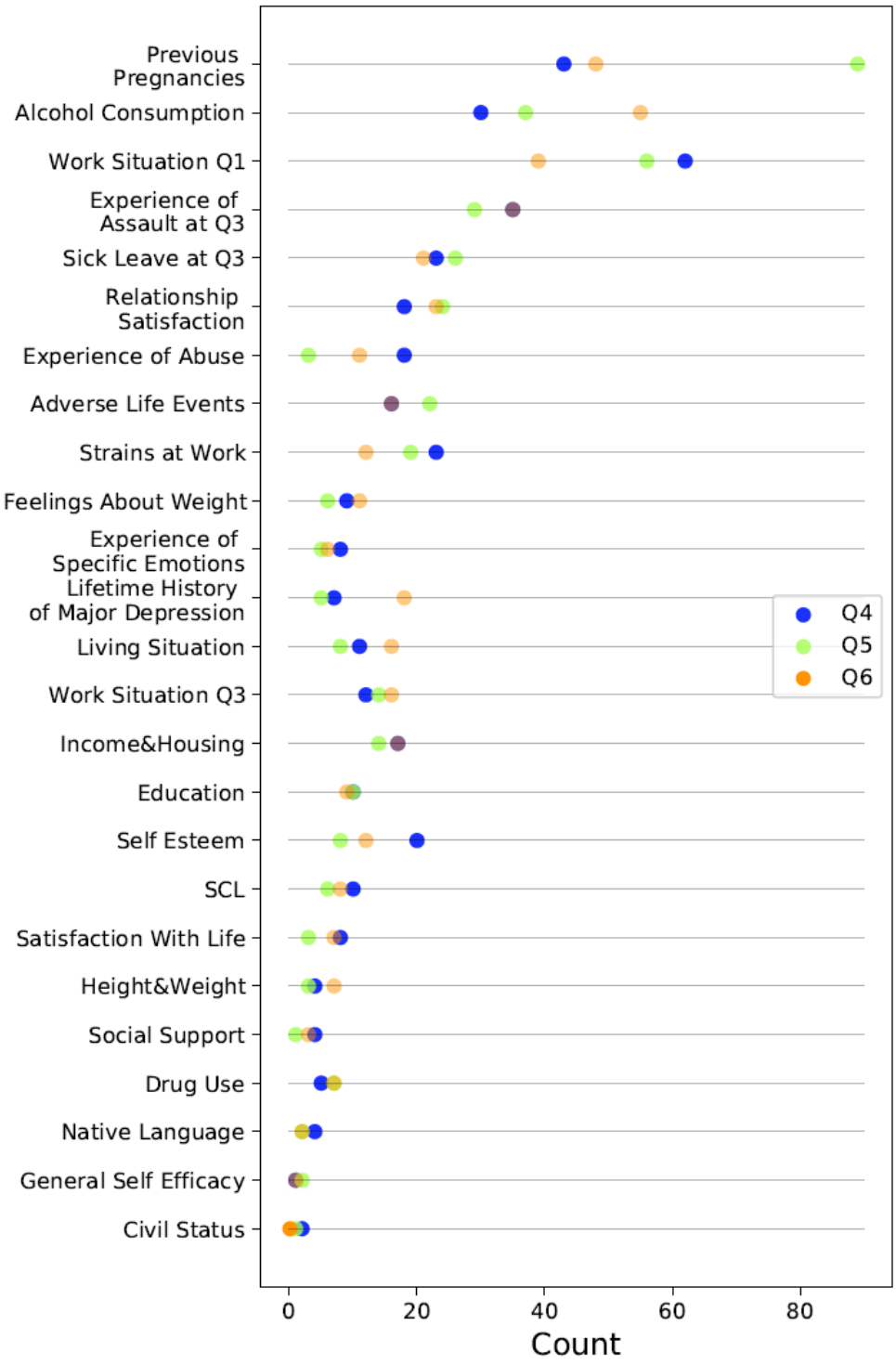

Figure 9.11: The figure displays the top 20 weights in each of the 20 principal components with the largest regression coefficients from an elastic net that predicted the mean SCL score at three time points. The prediction errors of the linear model can be found in Table 9.3.The group named "SCL" refers to the participants' answers to the SCL instrument at Q1 and Q3.

As in Figure 9.10, items related to alcohol consumption most frequently appear in the principal components with the largest regression coefficients in Figure 9.11. The frequency increases with time. The work situation at the beginning of the pregnancy experience a decline in frequency with time, which is natural given its relevance will become less important as time passes. The elastic net assigned higher regression coefficients to principal components where items related to previous pregnancies were more important than the linear regression model.

The groups of items with the highest importance from the XGBoost models are displayed in Figure 9.12.

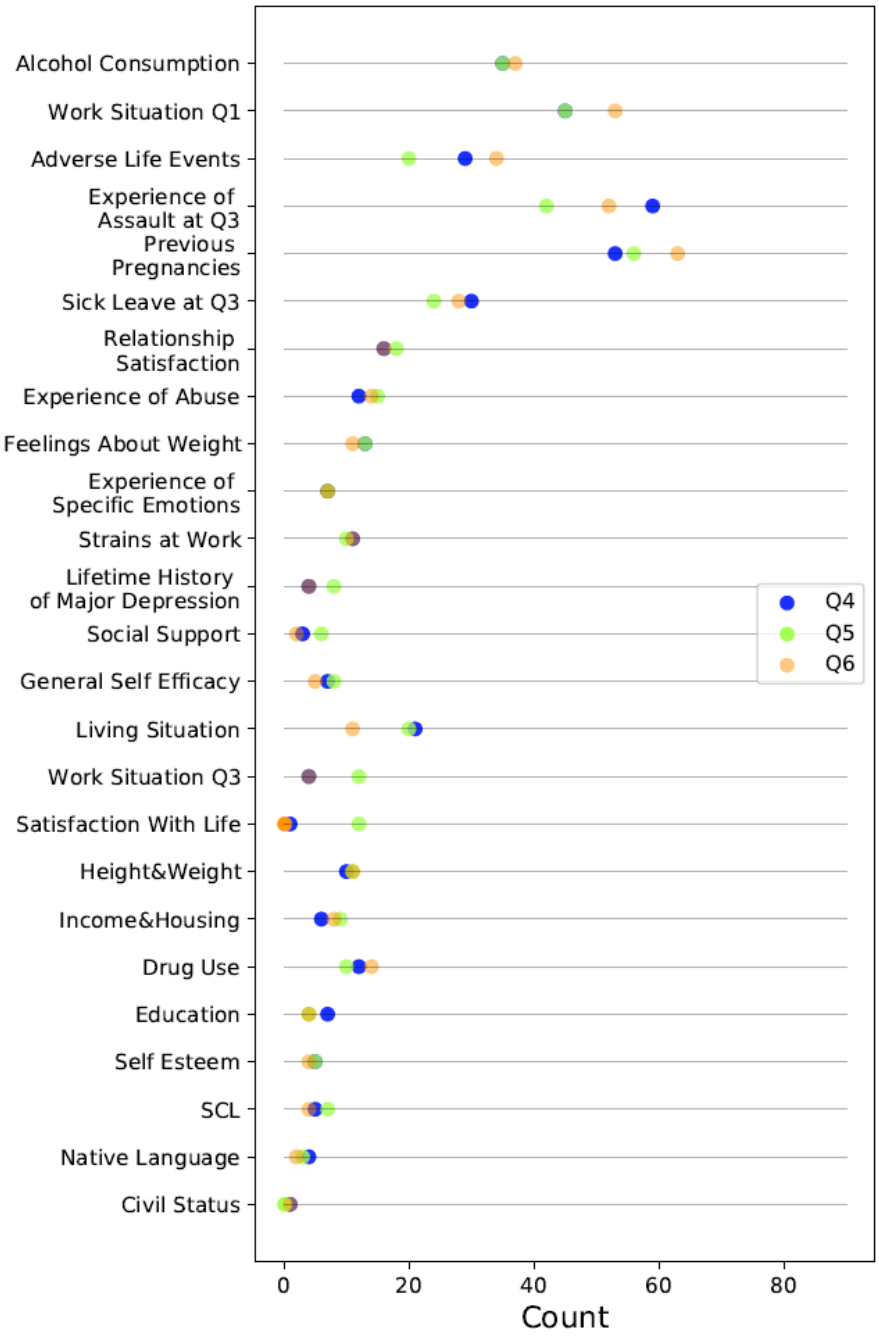

Figure 9.12: The figure displays the top 20 weights in each of the 20 principal components with the largest gain scores from an XGBoost model that predicted the mean SCL score at three time points. The prediction errors of the linear model can be found in Table 9.3. The group named "SCL" refers to the participants' answers to the SCL instrument at Q1 and Q3.

The two groups of items that appear on top of Figure 9.12 make it look like the prediction from Q4 is not included in the figure. However, the number of times items related to alcohol consumption and the work situation in Q1 appear in the principal components with the highest gain scores are equal for the models predicting the mean SCL scores in Q4 and Q5. We observe that the groups that appear most frequently are the same as in Figure 9.11.

9.2 Experiment 2: Investigating Predictive Ability and Feature Importance using Concurrent Exposures

9.2.1 Prediction Errors

The results associated with the second numerical experiment are presented. We predicted the mean SCL values using exposures measured concurrently at 6, 18 and 36 months postpartum (ref. research aim 1.2). A dataset for each specific time point was created. This differs from the previous experiment where the independent variables were collected during the prenatal period. The predictions made by the different models are presented in Table 9.9. By predicting

Table 9.9: Prediction errors for when the models where trained on the independent variables at each specific time point. The independent variables from Q1 and Q3 was not included.

Method	Q4	Q5	Q6	Mean
Multiple Linear Regression	0.3094	0.2632	0.3226	0.2984
Elastic Net	0.3054	0.2614	0.3188	0.2952
Neural Network	0.2995	0.2544	0.3090	0.2876
XGBoost	**0.2755**	**0.2370**	**0.2822**	**0.2649**

immediate mean SCL scores, as opposed to future scores, the models showed lower prediction errors. The XGBoost models had the lowest error. Compared to the neural network, the mean performance was 7.9% better and 11.2% better than the linear model.

9.2.2 Identifying Risk Factors in the Extended Postpartum Period

We studied the same importance measures as we did for the first experiment. However, in this experiment, we sorted the absolute value of importance measures from each model in descending order, i.e., the cluster with the highest importance has rank 1. The ranking of the different clusters of items is shown in Figure 9.13.

It becomes evident in Figure 9.13 that the linear regression models and elastic nets generally assign the same groups of items the same importance. At none of the time points do all three models agree on the rank of a specific group. The XGBoost models consistently rank items related to Rosenbergs self-esteem scale, smoking and the child's communication skills highly, while the linear regression models and the elastic nets rank items describing the experience of adverse life events and social support highly.

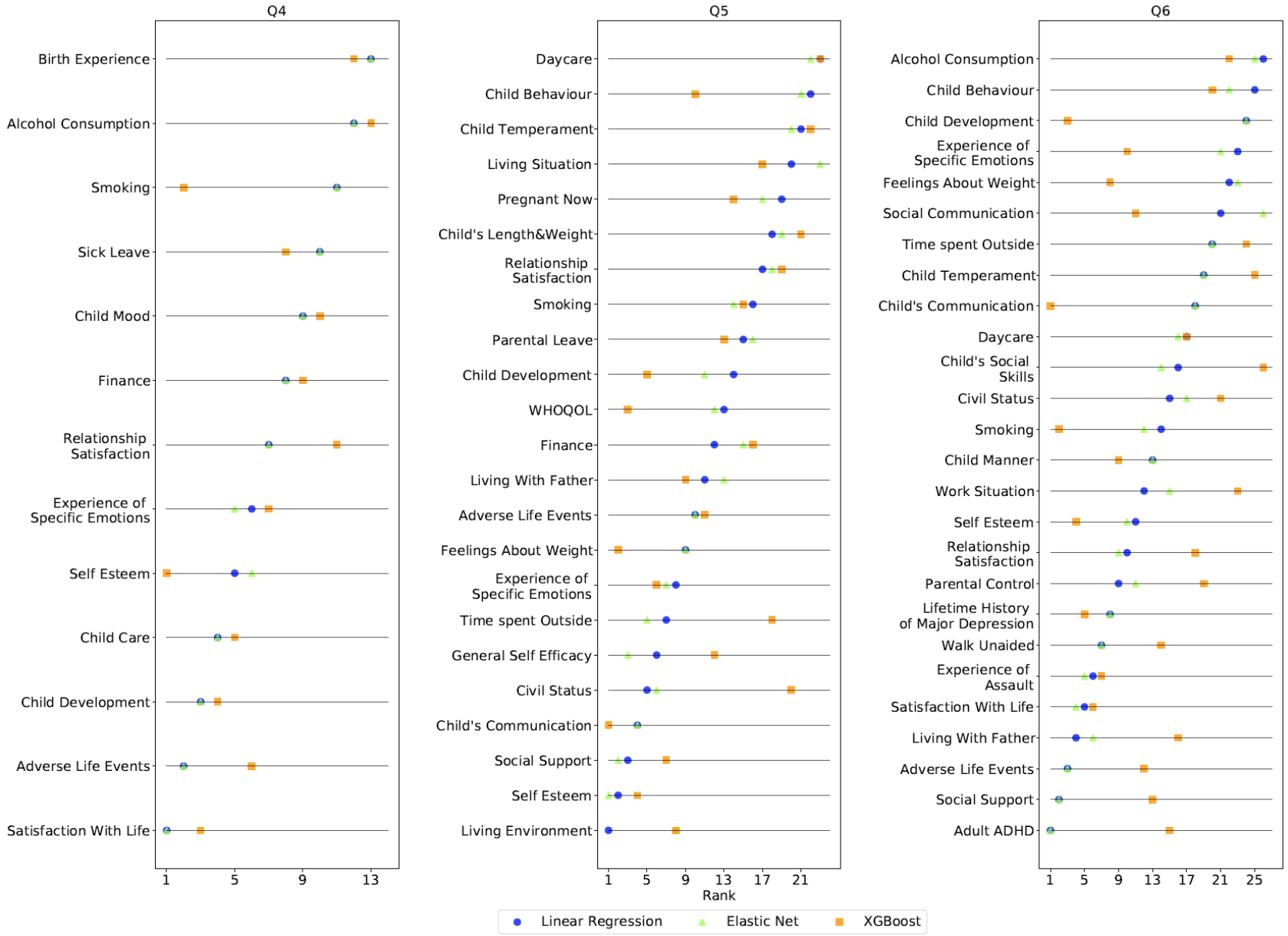

Figure 9.13: The ranking of the different clusters of items when predicting the mean SCL at Q4, Q5 and Q6 using the independent variables from each specific questionnaire. The ranking corresponds with the absolute magnitudes of the feature importance metrics. For the multiple linear regression models and elastics nets, the absolute value of the mean regression coefficients per group is calculated. For the XGBoost models, the gain scores are used as a metric. The group having the highest metric has rank 1, and the group with the lowest metric has the highest rank.

Chapter 10

Discussion and Limitations

10.1 Comparing Predictive Ability

In the first numerical experiment, where we predicted levels of anxiety and depression at 6, 18 and 36 months postpartum using features collected during the prenatal period on four different datasets, the neural networks and the XGBoost models alternated between producing the lowest prediction errors. As a reminder, the four datasets were created using i) aggregate scores on established scales, ii) item-level analyses, iii) dimensional reduction by principal component analysis, and iv) no dimensional reduction. We also created a fifth dataset consisting of features with an absolute correlation of 0.35 or higher with the outcome variable. The neural network showed the lowest error on the two data-driven datasets (iii and iv), while the XGBoost model produced the overall lowest prediction error on the correlated dataset, having a mean RMSE = 0.3716. The relative differences between the neural networks and XGBoost models were between 1% and 5%, with the largest observed difference on the correlated dataset. The prediction errors from the different models across all datasets revealed the following trend: the more sophisticated machine learning models outperformed the multiple linear regression model and the elastic net on all five datasets. The linear models showed consistently poorer predictive abilities and made errors between 6% and 8% higher than the best performing model at all times across datasets. Out of all the four methods, the multiple linear regression models had the worst predictive ability, which became especially evident in Figure 9.1.

The same trend was observed in the second numerical experiment. Here we predicted levels of anxiety and depression using exposures measured concurrently at 6, 18 and 36 months postpartum. The lowest prediction errors were obtained by the XGBoost model, with a mean RMSE = 0.2649. The mean RMSE of the XGBoost model was 7.9% lower than the one produced by the neural network and 11% lower than the mean error from the linear model. Compared to the first experiment, the prediction errors are overall lower when we are predicting at concurrent time points. This is not surprising. Several independent variables are expected to change with time. Thus, predicting future mean SCL scores with non-static independent variables is more challenging. This also explains why the lowest prediction errors were always obtained at Q4 in the first experiment.

10.2 Dimensionality and Performance

The predictions errors in the four tables 9.1 to 9.4 in subsection 9.1.1 shed light on some important aspects when it comes to predicting levels of depression and anxiety in the extended postpartum period. Aggregating features, instead of using the single items for the selected features, did not improve predictions. On the contrary, the predictions made with this dataset had the worst performance of all the datasets across all models. The mean of all the mean prediction errors for all models, measured using the RMSE, across time points was 0.3996 for the

aggregated feature dataset and 0.3934 when using the individual items. A possible explanation for these results can be that the aggregated features led to a loss of information, making it harder for the models to discriminate between different target values. This raises questions about how well we managed to construct informative features and the practice of aggregating features when high predictive abilities are desired for this case. We note that domain knowledge is paramount when creating new variables through aggregation, and how it is utilized can highly impact how these variables contribute to making meaningful predictions. There are many ways to combine features into new ones, and the subject of feature aggregation could make an interesting research topic on its own.

The use of aggregated features and feature engineering has proven fruitful both in research and online machine learning contests. If using the raw features always yielded the lowest prediction errors, all contestants would have identical results. This is not to say that our findings are invalid, but it highlights some of the aspects of the "no free lunch" principle in predictive modeling[159]. The principle states that there is no universally best algorithm for all problems. There could also exist more preferable combinations of the risk factors described in section 8.1 that would have produced lower prediction errors than the ones in Table 9.1. However, this is impossible to determine without further investigations.

Comparing the performance of the different models on the four original datasets revealed that all of the models, except for the XGBoost model, performed better on the two datasets created by taking a more agnostic approach to feature selection. The predictions made on the dataset constituted of the principal components yielded the lowest overall prediction errors, barely beating the dataset made up of all available items. The combined mean value across time points and methods was 0.3864 and 0.3878, respectively. A principal component analysis finds the directions within the space of observations that has the highest variance in order to identify relevant information. In the case where many of the original features are irrelevant to the target, one would expect that a model trained on the principal components would exhibit a higher predictive ability due to utilizing the more significant features repeatedly in the training procedure. From Figure 9.4 it is evident that over 50% of features had an absolute correlation coefficient below 0.2. However, the prediction errors from the subset of "highly" correlated features in Table 9.6 did not consistently produce lower errors for all models.

From the prediction errors made on the correlated subsets, there could be some indications of difficulties in discriminating between significant features, especially for the XGBoost models. The prediction errors revealed that the lowest errors obtained for all time points across all the different datasets belonged to the XGBoost model on the correlated subsets. From a naïve theoretical point of view, this behaviour can be seen as somewhat unexpected, given that the features used to create new nodes in a tree are the ones that yield the lowest gain score of all the proposed splitting points. However, when there are many features in the data, there is an increasing possibility that some combinations and interactions between weakly correlated features can affect the outcome even if there is no underlying relation. This phenomenon should be limited when applying regularization factors, such as controlling the depth of the trees. This could also indicate that the problem is an optimization problem and that more desirable hyperparameter combinations exist that our random search did not find. Nevertheless, pre-selecting features can seem to have an improving effect on the XGBoost models, given the low prediction errors obtained when using selected single items and highly correlated features.

An important point to emphasize is that the pairwise differences in prediction errors between the theory-based and data-driven datasets are small. We can not assert which dimensionality reduction method is the most desirable by only looking at prediction errors. It requires further testing, e.g., evaluation by cross-validation and a statistical significance test like a Student's t-test or calculating confidence intervals with bootstrapping. These results serve as a beacon of light for future research and highlight some important aspects of our original data and the choices we made when creating the different datasets. We argue that the prediction errors

ideally should be estimated from a resampling procedure, which would also give an estimate of the variance in prediction errors across different folds. Our intent from the start was to use cross-validation; however, due to inadequate computational resources and a technical issue related to exploding regression coefficients, the decision was made to test the models on a single holdout-set. Unfortunately, when we realized that we could not successfully perform the cross-validation procedure, time restraints inhibited us from collecting sufficient results through bootstrapping. In the future, we recommend using a resampling technique when producing predictions in order to easier discriminate between the different methods and datasets.

10.3 Identifying Clinical Levels of Depression and Anxiety

The distribution of predictions versus the real targets in Figure 9.2 revealed something that the prediction errors could not: how did the predictions fit the true underlying distribution. All of the models exhibited a peak in frequency around the mean values in the different test sets. On the surface, the linear model seemed to be the model that best captured the tails in the distributions, while the neural net had the largest amount of predictions falling close to mean values. Even though the elastic nets and the XGBoost models produced lower prediction errors than the multiple linear regression model and, on the surface, showed a greater ability to capture the tails in Figure 9.2, the residuals in Figure 9.3 revealed another reality: only a small proportion of the predictions were close to their target values. The lack of normality in the residuals and the fact that a large proportion of them was negative indicated that the predictions overestimated the target values. The same behaviour was not observed in the training set (ref. Figure 9.6), where all models showed residuals close to being normally distributed. One possible explanation for the overestimation could be the difference between $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. The splits were randomly created, and the training sets ended up having a larger proportion of participants with larger mean SCL scores.

The prediction errors from when the models were evaluated on the training set could partially explain the predictions made on the test set. Comparing the prediction errors made on the training and test sets can reveal much about the goodness of the fit for the model. Very low training errors compared to the test errors may indicate overfitting, while high errors on both sets can be a sign of underfitting. A good fit can be recognized by having a lower training error than test error, but the difference is usually small. The XGBoost models had the largest reduction in prediction errors, both in Table 9.7 and 9.8, of all the models. This could be a sign of overfitting, which would make generalization hard. On the other hand, it could be that the other models had more trouble identifying samples in the tail during training and that the training and test distributions are too different from each other, making generalization hard. The fact that the neural network had higher prediction errors on the training set than the test set could indicate underfitting and a high bias in the model. These problems could point to non-desirable hyperparameter combinations or train/test splits.

The number of individuals with high mean SCL scores in our dataset was relatively low, as revealed by Figure 8.2. A skewed target distribution should, in theory, not pose a problem. However, the low prevalence of higher values makes it difficult for the models to predict them correctly. One remedy to this problem is to transform the target variable in an effort to make it more normal. Common transformations for data with strictly positive values are either logarithm or square root transformations. They were both tried in different steps of our analysis process, but they did not improve our predictions. For this reason, we did not move forward with any of the transformations when we produced our final results. Another way of handling skewed targets is to alter the number of training samples, either by removing observations to balance out the training set or by resampling the more rarely observed data points. The former is known as undersampling, and the latter as oversampling. The terms are umbrella terms for several different techniques and are most commonly seen in classification problems. A naïve

undersampling approach was taken during the optimization process of the models by removing a proportion of the samples that had a mean SCL score below 1.85 (the cut-off score for the SCL-10). The alterations of the training set did not improve the out-of-sample errors, and we later discovered that this form of undersampling is not recommended for continuous targets. Therefore, a decision was made not to follow this path further.

The categorization of the predictions into binary values revealed another facet of the models' ability to capture the underlying nature of our sample. While all the models exhibited a high approximated specificity, all three machine learning models failed to correctly identify individuals with mean SCL scores above 1.85, which is the cut-off score for the SCL-10 instrument. Compared to the linear models, which had a mean approximated sensitivity of 56.0% when the models were trained on all available items, the machine learning methods had a mean sensitivity of less than 4%. This led us to raise questions about the representation of the relation between the target and independent variables in our data and the choice of using regression to identify individuals with clinical levels of anxiety and depression. Does a mapping function $\hat{f} : \mathbb{R}^N \to \mathbb{R}$ that could produce our observations even exists. If we instead had framed our initial problem as a binary classification problem, the predictions would reside within a much smaller space, which would arguably be an easier task than having an infinite amount of possible target values. It would also require a lower discriminatory ability from the models. However, categorizing the target variable would lead to a loss of information and the inability to explore non-clinical levels of depression and anxiety, which affect many women globally. The researcher would instead have to decide in advance which range of the instrument is of most interest.

In the future, if the use of machine learning models is to be used to help identify individuals at risk for developing clinical levels of depression and anxiety, the ability to capture the tail is critical. By taking an agnostic approach, we risked including insignificant variables in the training procedure that could lead the data to become noisy. This could explain why the models, especially the neural network, had a strong tendency to predict the mean value of the test set. Instead of random guessing, predicting the mean value would produce lower prediction errors on average. However, as Figure 9.5 showed, only including high correlated features in the training data did not improve the residuals and showed similar patterns as the ones observed in Figure 9.3.

10.4 Identifying Risk Factors

The previous analyses revealed that our models had problems capturing the individuals exhibiting the highest symptom levels of depression and anxiety in the extended postpartum period. So the features that were given the highest feature importance measures could be factors that better explain symptoms in the non-clinical range.

10.4.1 Prenatal Risk Factors

We used feature importance measures from the linear regression models, elastic nets and XGBoost models to train on all available features and the principal components making up 95% of the explained variance to identify prenatal risk factors. Regardless of which dataset was used as training data, some groups of features repeatedly exhibited high importance in all of the models.

One such group was alcohol consumption. Across the different models, the group had either noticeably higher regression coefficients or gain scores than the remaining groups. It was not possible to establish how the consumption of alcohol affected the mean SCL over time, as it showed varying importance across the different time points and models. The use of alcohol during pregnancy has been identified as a risk factor for experiencing depression and anxiety after giving birth [160], and women with untreated mental illness are more likely to consume

alcohol [161].

Substance abuse is also linked to depression and anxiety in mothers. While the gain scores for the items related to drug use were somewhat higher than the majority of the remaining groups, especially in Figure 9.9, they did not exhibit the same degree of importance as alcohol consumption. All of the items assessing substance abuse were mostly binary items, such as the ones given in Table 8.3, and they were highly imbalanced. Low variance in the independent variables can create difficulties when trying to model the relationship between them and the dependent variable. The low prevalence of participants with positive answers to the drug-related items can have several explanations: (i) attrition in longitudinal studies is inversely proportional to socioeconomic status. We exclusively included participants who had answered all six questionnaires. This may have created a biased sample. In 2004, 21.2% of all Norwegian women had completed a higher education program [162] and the estimated yearly salary was 254 476 NOK [163, p. 48]. Table 8.2 revealed that over 50% had higher education and an income above 200 000 in our sample. Our sample was higher educated than the general population, which also explains the high income level. (ii) The use of drugs is illegal, and thus giving positive answers can induce self-stigmatization and cognitive dissonance, restraining participants from giving truthful answers. (iii) Given the known health risks associated with drug use during pregnancy and its illegal nature, the prevalence is naturally lower than alcohol use.

In our grouping of single items, there were four groups related to the mother's work situation: strains at work, sick leave in Q3 and her work situation in Q1 and Q3. They all showed varying importance in the different models, and how they varied with time was somewhat inconsistent. In the elastic nets trained on all available items, the items related to the work situation in Q3 had the highest positive mean coefficient of all groups when predicting the mean SCL in Q4. It gradually became smaller with time, being almost zero when predicting the mean SCL at Q6. The items related to the work situation in Q1 experienced the opposite, with the mean coefficient for the group being almost zero in Q4 and gradually increasing with time. This behaviour is also observed in the linear models trained on the same data. The positive coefficients could indicate an association between higher mean SCL values and specific occupation groups. However, further testing would be needed to make more definitive claims, but on the surface, the results seem to be in line with the results from Clayborne et al. [43], which found that prenatal work stress was associated with both prenatal and postnatal depression and anxiety.

When the models were trained on all available features, the items related to how the participants felt about their weight and a potential weight gain during the pregnancy had the highest negative regression coefficients in the multiple linear model and the elastic net. They showed the highest gain score across all time points in the XGBoost model. The same importance was not shown when the models were trained on the principal components. Body image concerns have previously been shown to have a mediating effect on depressive symptoms and weight gain in the time after delivery [47]. Our models seem to capture some of the same underlying behaviour, and in the scenario where the findings in [47] were not yet discovered, the feature importance measures could serve as a guiding light for future research.

10.4.2 Risk Factors in the Extended Postpartum Period

Based on the rankings of the clusters of single items in Figure 9.13, it is evident that the multiple linear regression models and the elastic nets generally assigned the same importance to each group. The continuous variable selection in an elastic net did not seem to affect the final ranking of the groups. If we had investigated the magnitude of the different regression coefficients, we would expect them to be slightly lower than those from the multiple linear regression model. One possible explanation for the difference in how the models ranked the groups is the implicit interactions between the features in the XGBoost models. It is possible to model interactions with both the linear model and the elastic net. However, constructing meaningful interaction terms often requires domain knowledge and is a more manual process. A

possible improvement of our results could be obtained by implementing such interaction terms, and we propose this as an interesting topic to further investigate when comparing machine learning models with more conventional methods.

Making comparisons across time points in Figure 9.13 is not necessarily a reasonable approach, given the presence of different groups of features in each time point. One possibility is to filter out the items that appear in all three time points, redo the predictions and investigate the rankings. However, given the implicit interactions in the XGBoost model and the feature selection in the elastic net, the rankings would probably change, and we would lose information by dismissing features. One of the groups that appeared in all-time points was items related to Rosenberg's self-esteem scale. The XGBoost consistently ranked this group as important, while it was only ranked among the top three groups in Q5 by the two other models. Items related to the child's communication skills also received high rankings. Within this cluster, we find instruments assessing non-verbal communication, which is an autism screening tool. A study from 2019 [164] found that symptoms of autistic children were observed to be strongly associated with both maternal anxiety and depressive symptoms.

10.5 Interpretability and Explainability of the Sophisticated Models

In order to fully utilize the potential of machine learning models in the health care domain, we are dependent on the model's interpretability and explainability (ref. the discussion in subsection 2.3.2). The two more sophisticated models applied in this thesis, neural networks and gradient boosted regression trees, are both said to be black-box models, given their complexity and deep connections. In other words, they are not inherently interpretable nor explainable. The development of algorithms specifically designed to interpret predictions made by black boxes is a stand-alone research field[1], and the algorithms can either be model-agnostic or model-specific.

For the XGBoost model, there are several model-specific importance measures available, making it possible to identify which of the independent variables contributed the most to making accurate predictions. Many model-specific interpretation algorithms developed for neural networks are tailored towards convolutional neural networks and image recognition tasks, which can not necessarily be applied to tabular data. Other methods, such as Gedeon's method [165] and Garson's algorithm [166], use the network's weights to determine variable importance. However, in a deep neural network where the number of hidden layers exceeds two, the interpretation of weights becomes increasingly complex and intractable to interpret.

There are several model-agnostic algorithms that can be utilized with a neural network. Compared to the inherent importance measures that come with XGBoost, the agnostic interpretation algorithms often used to explain a network's predictions are expensive when it comes to computational power and time. An increasingly popular method for gaining insight into how the different independent variables affect the output in an arbitrary machine learning model is to calculate so-called SHAP values. These values are calculated based on a game-theoretic approach, and they represent a feature's responsibility for how the target value change [167]. An effort was made to calculate said values for our neural networks. However, we lacked the computational power to make the calculations feasible for the number of networks trained during our work. We recognize the limitation this has on our results, given that a substantial amount of effort was put into optimizing the networks for making accurate predictions while being unable to learn anything about how the predictions were formed.

The list of model agnostic algorithms developed in recent years is growing, e.g., see [84] for an extensive overview. However, the amount of work needed to implement them successfully

[1]Explainable artificial intelligence (XAI) is a research field with focus on developing tools and algorithms for interpreting, refining and validating machine learning models.

is not always trivial compared to the importance measures from XGBoost. We argue that this is one of the major drawbacks of neural networks, together with their sensitivity to different hyperparameter values and the computational cost required to obtain optimal combinations of said hyperparameters.

10.6 Recommendations

This thesis's third and final research aim is to present a set of recommendations on using machine learning to analyze registry and health survey data based on the experiences made from answering the first and second research aims. Based on the results and discussions above, we have comprised a list of what we deem to be reasonable recommendations.

- The sophisticated machine learning methods exhibited the best predictive abilities. Despite the neural networks exhibiting low prediction errors, the amount of time and work needed to adequately optimize them and interpret their predictions can be tedious compared to the XGBoost models. Thus, we recommend that researchers who want to explore machine learning and its potential on the MoBa data steer away from neural networks in favor of tree-based methods unless they desire to explore specific network architectures. This is in line with the field's general recommendation for tabular data[2]. However, there is a principle stating that there is no universally best machine learning algorithm for all problems, the so-called "no free lunch" principle [159], so it was not given in advance which of the four supervised learning models would be best suited for our task.

- Assuming that the aggregated features created from the selection of prenatal exposures were appropriately constructed, the errors in Table 9.1 and 9.2 indicated that the use of single items is preferred over aggregated features when predicting symptom levels of depression and anxiety. Hence, we recommend using single items over aggregated features.

- Using some form of feature selection can enhance the predictive abilities of the conventional linear regression model but also the machine learning methods. Three out of the four methods exhibited the lowest prediction errors when the data consisted of the principal components that made up 95% of the explained variance, and the XGBoost model had the best performance on a selection of "highly" correlated features. As with our first recommendation, this is already an incorporated procedure in the fields of machine learning and psychology. However, an important notice about using principal components is that they reduce the overall interpretability of the models.

[2] A recent study from November 2021 [168] found that extreme gradient boosting outperformed several deep learning architectures explicitly designed to fit tabulated data.

Chapter 11

Conclusions and Outlook

In this thesis, we have trained elastic nets, neural networks and gradient boosted trees through the XGBoost library on the Norwegian Mother, Father and Child Cohort study to predict levels of depression and anxiety in new mothers in the extended postpartum period and compared their predictive ability to multiple linear regression models. In psychology, the use of linear models is widespread, and one of the goals of this thesis was to compare the different models' predictive abilities on a novel dataset. By leveraging the supervised models' ability to learn non-linear relationships, we hypothesized that their predictive ability would exceed the one exhibited by the linear models and that we would be able to gain new insights into the underlying factors of depression afflicting mothers following childbirth. A long-term goal of applying machine learning to health data is the possibility of it contributing to identifying individuals at risk for developing depression, which would essentially ease the diagnostic procedure.

Our goals were formulated in three main research aims. The first one was directly concerned with different aspects of the predictive abilities of the different models, using exposures both from the prenatal- and extended postpartum period. Five different datasets were created from data collected during the prenatal period, two of them from a theory-driven approach where domain knowledge was applied. The remaining three took a more agnostic approach applying minimal to no selection criteria for the features. Our analyses found that the multiple linear models consistently showed higher RMSE values than the other models on all datasets. The errors ranged from 6% to 8% higher than the best performing model, and the XGBoost model achieved the best overall performance on the dataset with the features that had an absolute correlation of 0.35 or higher with the target variable. When the methods predicted symptom levels using exposures measured concurrently at 6, 18 and 36 months postpartum, the XGBoost model had the lowest mean prediction error. Compared to the neural network, the mean performance was 7.9% better and 11.2% better than the linear model.

When we evaluated the performance of different methods for identifying individuals at risk for clinical levels of depression and anxiety, our analyses revealed an overall poor ability to discriminate between high- and low-risk individuals for the machine learning models. The linear model had the highest approximated sensitivity of 56.0% when the models were trained on all available features, while the remaining models had sensitivity scores below 4%.

The second research aim was focused on identifying prenatal and concurrent exposures that can be of clinical interest or help inform theoretical models of post-party emotional problems. The feature importance measures from the linear models, elastic nets and XGBoost models were investigated to identify potential risk factors. This was done when the models were trained on prenatal and postnatal exposures. Three clusters of items appeared across all models when they were trained on prenatal items. Features related to alcohol consumption, attitudes about weight and weight gain during the pregnancy and items related to the mothers' working situation consistently showed high importance. When the postnatal items were used as training data, the models identified items related to Rosenberg's self-esteem scale and the child's non-verbal

communication skills as strong predictors for high symptoms levels.

Our third and final aim was to present a set of recommendations on using machine learning to analyze registry and health based on our experiences. When prediction is the primary goal, we recommended tree-based methods over neural networks due to less demanding tuning, training and post-analysis processes. We achieved lower prediction errors when working with single items compared to aggregated features, indicating one of two things: when applying machine learning to the MoBa data, the use of single items is preferred over aggregated items, or our new features were poorly constructed, leading to loss of information. Our final recommendation was to apply some form of feature selection process to the complete dataset, which is already considered a standard procedure in the field of machine learning and psychology.

We experienced several challenges with the more sophisticated models that the linear model did not experience. The encounter of non-normal residuals and low sensitivity prompted us to reconsider using regression as our strategy for predicting clinical levels of depression and anxiety in the extended postpartum period. Being able to clearly explain the models' predictions were more challenging for the sophisticated methods, especially the neural networks. Our work did not place us in a position where we could advocate for replacing the usage of linear models in favor of machine learning. They are inherently explainable and interpretable from the get-go, and they exhibited the highest ability to discriminate between individuals with higher mean SCL scores. However, they consistently produced higher prediction errors than the remaining models, so we are not disregarding the contributions from machine learning.

At the end of the day, the performance of any machine learning model is dependent on the data available and its quality. Before any modeling began, the data went through a comprehensive preprocessing procedure. The choices made when constructing the complete dataset may have affected how well our models were able to learn from it. Applying machine learning algorithms to novel datasets is ultimately a long process of trial and error, and in cases where the results do not live up to the expectations, there are still lessons to be learned from it. Through our work, we have established a baseline that can serve as a reference for future work. We argue that taking an agnostic approach through machine learning can provide valuable insights on research topics worthy of further investigation, both in the methodological domain and in the search for more profound etiological knowledge.

11.1 Outlook and Future Improvements

We describe some exciting topics that can be applied to our work to gain further insights into how machine learning can contribute to understanding the MoBa dataset in the future and other quantitative studies in epidemiology and fields in social sciences. Some new improvements are also described, all with the intent to enhance the work presented in this thesis.

We argued that the lack of discriminative ability for high-risk individuals found in the machine learning models could be due to our choice of predicting a continuous value instead of a binary outcome variable. A study similar to ours performed on longitudinal data from the U.S and South Korea found that deep-learning models detected depression better than conventional methods such as logistic regression [169]. The outcome variable was the binary status of depression. Framing our experiments as classification problems and redoing the analyses could be interesting for possibly improving the models' abilities to identify clinical levels of depression and anxiety. However, as we argued in subsection 8.3.1, categorizing a continuous variable can lead to loss of information and is not an accepted practice in all fields of research [21]. It has been shown that categorizing continuous variables can lead to a need for higher numbers of training data, with 100 observations of a continuous variable being statistically equivalent to at least 157 binary observations [170]. The decision to categorize the outcome variable has to be made with consideration, where each researcher has to decide what is most important in the specific case.

The imbalanced target distribution could also explain why the models had difficulties learning patterns that distinguished the high-risk individuals from the low-risk majority. Most of the work devoted to the topic of imbalanced regression is based on heuristically dividing the imbalanced dataset into two groups, infrequent and frequent observations, before applying classification-based methods [171]. By doing so, they fail to account for the distance between the targets and the similarity that occurs in variables that have roughly the same numeric value. In a recent article from 2021 on imbalanced regression problems, two algorithms incorporating this similarity both in label and feature space were developed [171]. Applying a specialized algorithm for skewed regression problems could potentially improve our models' discriminative abilities.

The results showed the highest prediction errors on the dataset containing aggregating features. Exploring new ways of combining the single items could result in us making different recommendations. The single items chosen for aggregation were selected based on domain knowledge. The lowest prediction errors were obtained on the dataset with the highest correlated features, indicating a need for a feature selection process. Exploring more sophisticated selection algorithms could prove fruitful and could be added to the preprocessing procedure.

The lack of interaction terms in the linear models and elastic nets could be a weakness of our models. We argued that constructing meaningful interaction terms required domain knowledge and could be time-consuming. However, this is only true if the procedure is started from scratch, i.e., starting from zero interaction terms and adding in one by one. Instead, a model consisting of all possible interaction terms could be created, and a recursive feature elimination process could take place to reduce the number of terms added to the model. It may be that the observed difference in prediction abilities could be reduced by including selected interaction terms.

During our discussion in chapter 10, we argued that we were not in a place to make bold claims regarding the method of choice and how to best handle the data to achieve low errors. The lack of definiteness in our argumentation is due to the results being obtained on a single hold-out set. We suggest collecting predictions through a k-fold cross-validation. If wanting to apply a significance test, we more specifically suggest a 5×2-fold cross-validation. For approximating statistical tests for supervised learning algorithms, the 5×2-fold cross-validation procedure was proposed by Dietterich [172]. The procedure repeats a regular cross-validation procedure five times, where both sets are randomly partitioned and have equal sizes. This procedure allows for directly measuring the variation due to the choice of training set [172]. Another possibility for establishing best-use methods is collecting bootstrapped predictions and using the bootstrap sample to calculate confidence intervals.

The second research aim focused on interpreting the models' predictions in order to identify prenatal or concurrent exposures that can be of clinical interest or help inform theoretical models of post-party emotional problems. It is a drawback that we could not retrieve any information from the neural networks and how they formed their predictions. We thus suggest implementing an agnostic interpretation algorithm that does not rely on layer weights in future experiments. As noted in chapter 10, this would require a substantial amount of work and could therefore serve as an interesting topic on its own.

One topic that we did not touch upon in chapter 10 was the time dependency in our data and how it was handled. The longitudinal nature of our dataset could be more explicitly incorporated into our choices of statistical models to achieve better predictive abilities. We chose the multiple linear regression model to represent the conventional model used in psychological research. However, as our discussion on state-of-the-art methods in section 2.3 revealed, when the data is longitudinal linear mixed effect models are often used (see Appendix A for theoretical background). An improvement is thus to predict levels of depression and anxiety with the linear mixed model, using time as a random effect. Mixed effect models have also recently been combined with boosting in an algorithm coined GPBoost [173]. Here the fixed effects are estimated using boosting and not a linear model. Recurrent neural networks are also used with longitudinal data, as they have the ability to preserve and exploit temporal relationships in the

data. They are designed for handling sequential data and have previously been applied in studies aiming at predicting diseases [174, 175]. Exploring any of these methods aimed at longitudinal data could possibly improve predictions and the possibility of incorporating machine learning in the screening process.

To successfully incorporate machine learning into the health care industry, we argue that the discriminative ability of the models has to experience some improvements. Our research found that the machine learning methods outperformed the linear model on predictive abilities but showed poorer ability to identify individuals with clinical symptom levels. The privacy regulations that often come with health data impose limitations on the progress attainable by machine learning. Reworking the dataset within an appropriate ethical and legal framework to potentially make it publicly available could excel the advancement of machine learning in the health care domain.

Part IV

Appendices

Appendix A

Additional Theory

This chapter includes theory on two subjects that we planned on incorporating in our work. However, as the work unfolded, we ran out of time, but the two subjects are still mentioned several places in our thesis. We first introduce hypothesis testing, followed by linear mixed models.

A.1 Hypothesis Testing

Making absolute statements about a specific population requires us to examine the *entire* population. This is impossible in most real-life scenarios, so we work with smaller population samples. Hypothesis testing, which is considered part of the field of inferential statistics, is a systematic way of making generalizations about a population from a smaller sample. When performing hypothesis testing, two hypotheses are created: the null- and alternative hypotheses. The former states that there is no statistical relationship to be inferred from the sample data, while the latter directly contradicts this statement and assumes a relationship. The desirable outcome is to reject the null hypothesis, which translates to there being a sufficient amount of evidence to support the alternative hypothesis. We always assume the null hypothesis to be true [176].

A.1.1 Statistical Tests and Level of Significance

To reject the null hypothesis, we must conduct a statistical test to determine if the probability of observing a statistical relationship in the observed data is higher than a pre-defined threshold. This threshold is often called the significance level or α-level and describes the probability of rejecting the null hypothesis when it is true. The field of psychology usually operates with the threshold of 0.05 [176]. Due to the significance level, we risk making two types of errors when forming conclusions from a statistical test: type I and type II errors. Type I error is to reject the null hypothesis when it is true, i. e. saying there is a statistical relationship when there is none. We make type II errors when we fail to reject the null hypothesis when it is false, meaning that we wrongfully assume that there is no statistical relationship.

The probability produced by the statistical test is called the p-value and describes the probability of observing a statistical relationship if the null hypothesis is true. The results are statistically significant if the computed p-value is lower than the α-level, indicating that the observed relationship has a probability of less than 5% of being due to chance.

A.1.2 Paired Student's t-Test

Hypothesis testing can be applied to determine a significant difference between the linear models and the more sophisticated machine learning models and their ability to learn patterns from the

MoBa dataset. The null hypothesis would then state that there is no difference between the different algorithms' learning abilities. The Student's t-test compares the means of two groups, assuming that the underlying distribution is normally distributed [177]. More formally, the null hypothesis states that the difference between the two group means is zero.

To determine if the means differ form one another, the t statistic is calculated, where large values of t indicate a significant difference between the groups. The t statistic is given by

$$t = \frac{\mu_1 - \mu_2}{s_p\sqrt{(N_1^{-1} + N_2^{-1})}}, \tag{A.1}$$

where s_p is an estimator of the squared pooled variance, also called the pooled standard deviation, μ_i and N_i are the group mean and sample size for the two groups with $i = 1, 2$. The pooled standard deviation is a measurement of dispersion for several populations when the mean value may vary between the populations, but the variance is assumed to be the same across all populations. In the case where the number of populations is m and the sample sizes are equal across the m groups, the estimator for s_p^2 is defined as

$$s_p^2 = \frac{\sum_{i=1}^{m}(n-1)s_i^2}{\sum_{i=1}^{m} n}, \tag{A.2}$$

where s_i^2 is the sample variance

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y}_i)^2. \tag{A.3}$$

Here y is the dependent variable, and the bar refers to the sample mean. The t value follows the Student's t-distribution, which resembles the normal distribution but has heavier tails, meaning that it is more likely to yield values far from the mean. To determine the p-value from the t statistic, one must consult a t-distribution table.

A.2 Linear Mixed Models

Linear mixed models (LMMs) are an extension of multiple linear regression models and are particularly useful when repeated measurements are made on the same unit of observation. Here the data will be correlated and violate the assumptions about linear independence in predictors from sec. 5.1. LMMs require data to be stored in the long format, given that each unit of observation now has several time-dependent response variables and features. For every ith unit of observation, at time point j, the input data is on the form $(x_{i1}^j, x_{i2}^j, ..., x_{ip}^j, y_{ij})$ for $i = 1, ..., N$ and $j = 1, ..., n_i$, with n_i being total number of time points for unit i and p is the total number of features. Each response variable is thus a vector $\boldsymbol{y}_i \in \mathbb{R}^{n_i}$ and the features are stored in a $(N \cdot n_i) \times p$ matrix denoted as $\mathcal{X}_i$.

LMMs captures both population- and individual effects, hence the design matrix $\mathcal{X}_i$ must be partitioned into two matrices: one storing features considered to be fixed, X_i, and one storing the random features, Z_i. In matrix notation, the LMM describes the response through the relationship

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{\nu}_i + \boldsymbol{\varepsilon}_i, \tag{A.4}$$

where $\boldsymbol{\nu}_i \sim \mathcal{N}(0, G) \in \mathbb{R}^p$ is an unknown random effect vector assumed to be normally distributed with zero mean and covariance matrix G, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, R) \in \mathbb{R}^{n_i}$ and $\boldsymbol{\beta}$ being the fixed effect regression coefficients [178]. The random effects are considered to be deviations from the regression coefficients.

A.2.1 Estimation Methods

To simplify the notation, we omit the subscript i, indicating that we are looking at the response for unit i in the rest of this section. The model in eq. (A.4) can be recast into the linear model,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}*, \tag{A.5}$$

with $\boldsymbol{\varepsilon}* \equiv Z\boldsymbol{\nu} + \boldsymbol{\varepsilon}$. The new variable $\boldsymbol{\varepsilon}*$ have the following expecation value and covariance matrix:

$$\mathbb{E}(\boldsymbol{\varepsilon}*) = Z\mathbb{E}(\boldsymbol{\nu}) + \mathbb{E}(\boldsymbol{\varepsilon}) = 0, \tag{A.6}$$

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\varepsilon}*) &= \mathrm{Cov}(Z\boldsymbol{\nu}) + \mathrm{Cov}(\boldsymbol{\varepsilon}),\\ &= Z\mathbb{E}[\boldsymbol{\nu}\boldsymbol{\nu}^T]Z^T = ZGZ^T + R.\end{aligned} \tag{A.7}$$

Meaning that $\boldsymbol{\varepsilon}* \sim \mathcal{N}(0, ZGZ^T + R)$, making $\boldsymbol{y} \sim \mathcal{N}(X\boldsymbol{\beta}, ZGZ^T + R)$. How to estimate the model in eq. (A.5) depends on wether G and R are known or not, we will cover both cases beginning with the former scenario.

Known Covariance Matrices

We will first describe how to estimate the fixed parameters $\boldsymbol{\beta}$, followed by the theory on how to predict the random effects $\boldsymbol{\nu}$.

When the covariance matrix $ZGZ^T + R \equiv V$ is known, we can estimate $\boldsymbol{\beta}$ using GLS. Hence, the fixed effect coefficients are given by eq. (5.10). The random effects $\boldsymbol{\nu}$ are usually predicted through the conditional expectation value $\mathbb{E}(\nu|\boldsymbol{y})$. We note that $\boldsymbol{\nu}$ and $\boldsymbol{y}$ are jointly normal and can thus be expressed as the bivariate normal distribution vector

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{\nu}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}X\boldsymbol{\beta}\\ 0\end{pmatrix}, \begin{pmatrix}V & ZG\\ GZ^T & G\end{pmatrix}\right), \tag{A.8}$$

with ZG being $\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{\nu})$ and $GZ^T = \mathrm{Cov}(\boldsymbol{\nu}, \boldsymbol{y})$. The bivariate conditional expectation value $\mathbb{E}(\nu|\boldsymbol{y})$, as expressed in eq. (3.4), and thus the predicted random effects $\hat{\boldsymbol{\nu}}$, is

$$\hat{\boldsymbol{\nu}} = GZ^T V^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}). \tag{A.9}$$

Unknown Covariance Matrices

When the covariance matrices are unknown, we assume that G and R are dependent on a vector of parameters $\boldsymbol{\vartheta}$, hence making V a function of $\boldsymbol{\vartheta}$,

$$V(\boldsymbol{\vartheta}) = ZG(\boldsymbol{\vartheta})Z^T + R(\boldsymbol{\vartheta}). \tag{A.10}$$

Estimating the fixed and random effects requires knowledge about V, which is no longer available. The solution: estimate V by estimating the parameter vector $\boldsymbol{\vartheta}$. If we are able to estimate $\boldsymbol{\vartheta}$ such that $V(\hat{\boldsymbol{\vartheta}})$ can act as the real value of V, we can estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\nu}}$ using eq. (5.10) and eq. (A.9), respectively.

The parameter vector $\boldsymbol{\vartheta}$ is estimated with MLE. The joint probability distribution of $\boldsymbol{y} \sim \mathcal{N}(X\boldsymbol{\beta}, V)$ is

$$p(\boldsymbol{y}) = (2\pi)^{-n/2}(\det V(\boldsymbol{\vartheta}))^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V(\boldsymbol{\vartheta})^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right), \tag{A.11}$$

with the log-likehood being

$$l(\boldsymbol{\beta}, \boldsymbol{\vartheta}) = -\frac{1}{2}\big(\log(2\pi) + \log(\det V(\boldsymbol{\vartheta})) + (\boldsymbol{y} - X\boldsymbol{\beta})^T V(\boldsymbol{\vartheta})^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\big). \tag{A.12}$$

Here $\boldsymbol{\beta}$ is a nuisance parameter, and thus the log-likelihood function in eq. (A.12) is maximized through the profile log-likelihood. As we recall from sec. 3.2.1, the profile log-likelihood is obtained by assuming that $\boldsymbol{\vartheta}$ is known, and maximize eq. (A.12) with respect to $\boldsymbol{\beta}$, holding $\boldsymbol{\vartheta}$ fixed, i. e.

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}) = \arg\max_{\beta} l_{\vartheta}(\boldsymbol{\beta}).$$

We recognize the maximized $\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta})$ as the GLS solution retrieved from applying the MLE formalism to the linear regression problem from section 5.1. The profile log-likehood function can be expressed as

$$l_p(\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}), \boldsymbol{\vartheta}) = -\frac{1}{2}\Big(\log(\det V(\boldsymbol{\vartheta})) + (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}))^T V(\boldsymbol{\vartheta})^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}))\Big). \tag{A.13}$$

Finally the estimated $\hat{\boldsymbol{\vartheta}}$ is obtained from maximizing l_p with respect to $\boldsymbol{\vartheta}$,

$$\begin{aligned} \hat{\boldsymbol{\vartheta}} &= \arg\max_{\vartheta} l_p, && \text{(A.14)} \\ &= \arg\max_{\vartheta} l(\arg\max_{\beta} l_{\vartheta}(\beta), \boldsymbol{\vartheta}). \end{aligned}$$

(A.15)

To sumarize, we estimate the fixed and random effects when the covariance matrices R and G are unknown from the same equations as when they are known, but substituting V with $V(\hat{\boldsymbol{\vartheta}})$,

$$\hat{\boldsymbol{\beta}}_{\text{LMM}} = (X^T V(\hat{\boldsymbol{\vartheta}})^{-1} X)^{-1} X^T V(\hat{\boldsymbol{\vartheta}})^{-1} \boldsymbol{y}, \tag{A.16}$$

$$\hat{\nu}_{\text{LMM}} = G(\hat{\boldsymbol{\vartheta}}) Z^T V(\hat{\boldsymbol{\vartheta}})^{-1} (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{\text{LMM}}). \tag{A.17}$$

Appendix B

Bias-Variance Decomposition for Quadratic Loss

In subsection 2.3.3 we describe the expected generalization error a model makes when using a quadratic loss in regards to the bias-variance tradeoff. Here the quadratic loss is the mean squared error (MSE) described in section 4.3. The MSE is given as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(\boldsymbol{x}^{(i)}))^2. \qquad \text{(eq. (4.15) revisited)}$$

where $\hat{f}(\boldsymbol{x}^{(i)})$ is our predictions for a input vector $\boldsymbol{x}^{(i)} = (x_1, \ldots, x_N) \in \mathbb{R}^N$, and y_i is given by

$$y_i = f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}_{\text{true}}) + \varepsilon_i \qquad \text{for } i = 1, ...N. \qquad \text{(eq. (4.1) revisited)}$$

Here $\boldsymbol{\theta}_{\text{true}}$ is the assumed true parameters needed to describing our observations, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is independent and identically distributed noise with zero mean and standard deviation σ. The expected generalization error can be decomposed into, in vector notation,

$$\mathbb{E}[(\boldsymbol{y} - \boldsymbol{f})^2] = (\boldsymbol{f} - \mathbb{E}[\hat{\boldsymbol{f}}])^2 + \mathbb{E}\left[(\mathbb{E}[\hat{\boldsymbol{f}}] - \hat{\boldsymbol{f}})^2\right] + \boldsymbol{\sigma}^2, \tag{B.1}$$

with $\hat{\boldsymbol{f}} \equiv \hat{f}(X)$ and $\boldsymbol{f} \equiv f(X)$, and X being the matrix storing the observed data. The first term in eq. (B.1) is the squared bias of $\hat{f}$, followed by the variance and the last term is the variance of the noise.

To arrive at eq. (B.1) we start from

$$\begin{aligned}\mathbb{E}\left[(y - \hat{f})^2\right] &= \mathbb{E}\left[(f + \varepsilon - \hat{f})^2\right], \\ &= \mathbb{E}\left[(f + \varepsilon - \hat{f} + \mathbb{E}[\hat{f}] - \mathbb{E}[\hat{f}])^2\right].\end{aligned}$$

where we have omitted the bold vector notation for convenience. Applying a good deal of algebra to the expression inside the brackets we end up with the relation

$$\begin{aligned}\mathbb{E}\left[(y - \hat{f})^2\right] &= \mathbb{E}\Big[(f - \mathbb{E}[\hat{f}])^2 + \varepsilon^2 + (\mathbb{E}[\hat{f}] - \hat{f})^2 + 2(f - \mathbb{E}[\hat{f}])\varepsilon + 2(\mathbb{E}[\hat{f}] - \hat{f})\varepsilon + 2(\mathbb{E}[\hat{f}] - \hat{f})(f - \mathbb{E}[\hat{f}])\Big], \\ &= \mathbb{E}\left[(f - \mathbb{E}[\hat{f}])^2\right] + \mathbb{E}\left[\varepsilon^2\right] + \mathbb{E}\left[(\mathbb{E}[\hat{f}] - \hat{f})^2\right], \\ &\qquad + 2(f - \mathbb{E}[\hat{f}])\underbrace{\mathbb{E}[\varepsilon]}_{=0} + 2\mathbb{E}[\mathbb{E}[\hat{f}] - \hat{f}]\underbrace{\mathbb{E}[\varepsilon]}_{=0} + 2\underbrace{\mathbb{E}[\mathbb{E}[\hat{f}] - \hat{f}]}_{=0}(f - \mathbb{E}[\hat{f}]), \\ &= \mathbb{E}\left[(f - \mathbb{E}[\hat{f}])^2\right] + \mathbb{E}\left[(\hat{f} - \mathbb{E}[f])^2\right] + \mathbb{E}[\varepsilon^2].\end{aligned}$$

The expectation values with respect to the dataset and noise are assumed to be independent. We have now showed how the expected prediction error can be decomposed into a bias, variance and error term, which highlights some the challenges associated with the bias-variance tradeoff.

Appendix C

Longitudinal Studies

The MoBa study is a cohort study, a form of a longitudinal study design where the participants share a common characteristic or experience and are followed over a period of time, with repeated measurements of selected characteristics or exposures. The design is well suited for studying individual change over time, evaluating the relationship between risk factors and the development of disease and treatment outcomes [179]. In general, longitudinal datasets consists of a dependent, or outcome, variable y_{it} and a feature vector $\boldsymbol{x}^{(i)} \in \mathbb{R}^p$, measured at time points $t = 1, ..., n_i$ for participant $i = 1, ..., N$, with N being the size of the participant sample. Note that the number of time measurements can differ between subjects. Features are also called predictors or independent variables. We will use these terms interchangeably throughout this thesis.

There are several different longitudinal study designs, with the majority being observational studies. Two popular observation designs are repeated cross-sectional studies, where the participant sample differs across time, and prospective studies, where the same participant sample is followed for the whole study duration. Cohort studies are thus a prospective study design.

We include a side note on different data formats used with longitudinal data. Each feature vector $\boldsymbol{x}^{(i)}$ can be considered a row vector in what is called a design matrix X, so that the matrix element X_{ij} is the value of the jth independent variable for the ith participant. Data storage in a design matrix is referred to as a wide data storage format. Another approach is to store each time point for each participant in a separate row, with the time-invariant variables appearing in each row. This is referred to as a long data storage format. A key difference between the two formats is that time is an explicit predictor in the long format, while it appears as an implicit predictor in the wide format. The raw data from MoBa is stored in a wide format.

C.0.1 Challenges with Longitudinal Study Designs

As with any study design, a longitudinal design has strengths and weaknesses. One significant advantage is studying individual trajectories of development and identifying and relating events to specific exposures. On the other hand, it is normal for longitudinal studies to experience some loss in the participant sample during the study; this is called attrition or drop-out. Studies with more extended follow-up periods experience higher attrition rates, which can lead to a biased sample. Known risk factors for drop-out are low educational level, unemployment, being unmarried and an unhealthy lifestyle [180]. Attrition reduces the overall variability of the data and leads to missing data points in the dataset.

Missing data is a concern in and of itself, and it may be one of the most challenging issues to handle since there is no consensus on how to resolve the problem. Missing data is usually a result of attrition or unanswered questions in longitudinal studies. Improper handling of this data might have a significant negative impact on the statistical analyses. Some analyses simply

exclude participants with missing data, which leads to a reduction in the sample size. The reduction may jeopardize the statistical power of the analyses. Here statistical power refers to the model's ability to reject the hypothesis that no statistical relationship exists between the dependent and independent variables, also called the null hypothesis. One way to circumvent this problem is by imputing the data. Imputation is to apply a statistical procedure to replace missing data points and is a topic for further discussion in chapter 4, along with missing data mechanisms.

Correlation between repeated measurements on the same participant can also be problematic, as longitudinal studies are correlated by design. A measurement sampled at time t_1 will be correlated with the subsequent measurement at time t_2. This is called intraindividual correlation. Suppose the participants are sampled across different sites, e.g., geographical locations, correlation can arise within the sites due to human bias imposed by the research conductors, study protocol variations or inconsistencies in equipment across sites. Unequal sample sizes can also be challenging and lead to difficulty in detecting effects when participants are sampled at different sites [71].

Appendix D

Single Item Overview

For replication purposes we list the all of the item IDs that were used to construct the different datasets. The prefix in each ID correspond with which questionnaire the item belonged to, e.g., "AA" refers to Q1 and "CC" to Q3.

D.1 Applying Domain Knowledge: Selected Prenatal Features

We created two datasets from selected risk factors during the prenatal period: one where the single items where aggregated into new features, and one without any aggregation. All of the items are listed in Table D.1.

Table D.1: All unique item IDs for the single items that made up the subsets from the feature selection process.

Selected Item IDs from Q1 and Q3								
AA1548	AA1549	AA1550	AA1551	AA1552	CC1202	CC1203	CC1204	CC1205
CC1206	CC1207	CC1208	CC1209	AA1532	AA1533	AA1534	AA1535	AA1536
AA1537	AA1538	AA1539	AA1540	AA1541	CC1192	CC1193	CC1194	CC1195
CC1196	CC1197	CC1198	CC1199	CC1200	CC1201	CC1251	CC1252	CC1253
CC1258	CC1259	CC1260	CC1265	CC1266	CC1267	CC1272	CC1273	CC1274
CC1216	CC1217	CC1218	CC1219	CC1220	CC1213	CC1214	CC1215	CC1229
CC1230	CC1231	CC1232	CC1233	CC1235	CC1237	CC1239	CC1241	CC1243
CC1245	CC1247	CC1249	AA1572	AA1577	AA1579	AA1545	AA1546	AA1547
AA1124	AA1125	AA1305	AA1315	AA1316	AA1527	AA1528	AA1529	AA1530
AA1531								

D.2 All Available Prenatal Features

All of the available prenatal features are listed in Table D.2.

D.3 Postnatal Features

We created three different datasets containing all available features from Q4, Q5 and Q6. All features from Q4 are listed in Table D.3, while Table D.4 and Table D.5 list the items from Q5 and Q6, respectively.

Table D.2: Item IDs belonging to all available features from Q1 and Q3 (the prenatal period).

All Available Item IDs from Q1 and Q3													
AA1172	AA1173	AA1174	AA1175	AA1176	AA1177	AA1178	CC923	CC924	CC925	CC926	CC927	CC928	CC929
AA1452	AA1453	AA1454	AA1455	AA1456	AA1457	AA1458	AA1459	AA1460	AA1461	AA1462	AA1463	AA1464	AA1465
AA1466	AA1467	AA1468	AA1469	AA1470	AA1471	AA1472	AA1473	AA1474	CC1156	CC1157	CC1158	CC1159	CC1160
CC1161	CC1162	CC1163	CC1164	CC1165	CC1166	CC1167	CC1168	CC1169	CC1170	CC1171	CC1172	CC1173	CC1174
CC1175	CC1176	CC1177	AA1527	AA1528	AA1529	AA1530	AA1531	CC1224	CC1225	CC1226	CC1227	CC1228	AA1532
AA1533	AA1534	AA1535	AA1536	AA1537	AA1538	AA1539	AA1540	AA1541	CC1192	CC1193	CC1194	CC1195	CC1196
CC1197	CC1198	CC1199	CC1200	CC1201	AA1545	AA1546	AA1547	CC1179	CC1180	CC1181	AA1548	AA1549	AA1550
AA1551	AA1552	CC1202	CC1203	CC1204	CC1205	CC1206	CC1207	CC1208	CC1209	AA1568	AA1569	AA1570	AA1571
CC1229	CC1230	CC1231	CC1232	CC909	CC910	CC911	CC912	CC913	CC914	CC915	CC917	CC918	CC919
CC920	CC921	CC922	CC935	CC936	CC937	CC940	CC941	CC942	CC943	CC944	CC945	CC946	CC948
CC949	CC950	CC951	CC952	CC953	CC954	CC956	CC957	CC958	CC959	CC960	CC961	CC962	CC964
CC965	CC966	CC967	CC968	CC969	CC970	CC1067	CC1068	CC1069	CC1070	CC1071	CC1072	CC1073	CC1074
AA1123	CC1178	CC1210	CC1211	CC1212	CC1213	CC1214	CC1215	CC1216	CC1217	CC1218	CC1219	CC1220	CC1233
CC1234	CC1235	CC1236	CC1237	CC1238	CC1239	CC1240	CC1241	CC1242	CC1243	CC1244	CC1245	CC1246	CC1247
CC1248	CC1250	CC1251	CC1252	CC1253	CC1254	CC1255	CC1256	CC1257	CC1258	CC1259	CC1260	CC1261	CC1262
CC1263	CC1264	CC1265	CC1266	CC1267	CC1268	CC1269	CC1270	CC1271	CC1272	CC1273	CC1274	CC1275	CC1276
CC1277	CC1278	AA85	AA86	AA87	AA88	AA89	AA93	AA94	AA95	AA96	AA97	AA98	AA99
AA100	AA101	AA102	AA103	AA104	AA105	AA106	AA107	AA108	AA109	AA110	AA111	AA112	AA113
AA114	AA115	AA116	AA117	AA118	AA119	AA120	AA121	AA122	AA123	AA124	AA125	AA126	AA127
AA128	AA129	AA130	AA131	AA132	AA133	AA134	AA135	AA136	AA137	AA138	AA139	AA140	AA141
AA142	AA143	AA144	AA145	AA146	AA147	AA148	AA149	AA150	AA151	AA152	AA153	AA154	AA155
AA156	AA157	AA158	AA159	AA160	AA161	AA162	AA163	AA1124	AA1125	AA1126	AA1127	AA1128	AA1129
AA1130	AA1131	AA1179	AA1180	AA1181	AA1182	AA1183	AA1184	AA1185	AA1186	AA1187	AA1188	AA1189	AA1190
AA1191	AA1192	AA1193	AA1194	AA1132	AA1133	AA1134	AA1135	AA1136	AA1137	AA1138	AA1139	AA1140	AA1141
AA1142	AA1143	AA1144	AA1145	AA1146	AA1147	AA1148	AA1149	AA1150	AA1151	AA1152	AA1153	AA1157	AA1159
AA1160	AA1161	AA1162	AA1163	AA1164	AA1166	AA1167	AA1293	AA1294	AA1295	AA1296	AA1297	AA1298	AA1300
AA1301	AA1302	AA1303	AA1304	AA1305	AA1309	AA1315	AA1316	AA1317	AA1318	AA1319	AA1320	AA1321	AA1322
AA1323	AA1324	AA1325	AA1326	AA1327	AA1328	AA1475	AA1476	AA1477	AA1478	AA1479	AA1480	AA1481	AA1482
AA1483	AA1484	AA1485	AA1486	AA1487	AA1488	AA1553	AA1554	AA1555	AA1556	AA1557	AA1558	AA1559	AA1560
AA1561	AA1562	AA1563	AA1564	AA1565	AA1566	AA1567	AA1572	AA1573	AA1574	AA1575	AA1576	AA1577	AA1578
AA1579													

Table D.3: All unique item IDs for the available items in Q4.

All Available Item IDs from Q4									
DD774	DD775	DD776	DD777	DD778	DD779	DD800	DD801	DD802	DD803
DD804	DD784	DD785	DD786	DD787	DD788	DD789	DD790	DD791	DD792
DD793	DD837	DD838	DD839	DD840	DD841	DD842	DD843	DD844	DD833
DD834	DD835	DD836	DD746	DD747	DD748	DD749	DD750	DD751	DD752
DD753	DD754	DD755	DD756	DD757	DD758	DD759	DD760	DD765	DD766
DD767	DD768	DD769	DD770	DD771	DD772	DD773	DD794	DD795	DD796
DD797	DD798	DD799	DD805	DD806	DD807	DD808	DD809	DD810	DD811
DD812	DD813	DD814	DD815	DD816	DD817	DD818	DD819	DD820	DD821
DD822	DD823	DD824	DD826	DD12	DD13	DD14	DD15	DD16	DD17
DD18	DD20	DD21	DD22	DD23	DD24	DD25	DD26	DD27	DD28
DD30	DD33	DD34	DD35	DD36	DD37	DD38	DD39	DD40	DD41
DD348	DD349	DD350	DD351	DD352	DD353	DD354	DD355	DD356	DD357
DD358	DD359	DD360	DD361	DD362	DD363	DD364	DD365	DD366	DD367
DD386	DD387	DD388	DD389	DD390	DD391	DD392	DD393	DD394	DD395
DD674	DD676	DD677	DD678	DD679	DD680	DD682	DD683	DD684	DD685
DD686	DD688	DD689	DD690	DD691	DD692	DD693	DD694	DD732	DD733
DD734	DD735	DD736	DD737	DD738	DD739	DD740	DD741	DD742	DD743
DD1114	DD1115	DD1116	DD1117	DD1118	DD1119	DD744	DD745	DD827	DD828
DD829	DD830	DD831	DD832						

Table D.4: All unique item IDs for the available items in Q5.

All Available Item IDs from Q5												
EE607	EE608	EE609	EE925	EE926	EE927	EE928	EE929	EE930	EE931	EE932	EE933	EE934
EE935	EE936	EE937	EE938	EE939	EE940	EE941	EE942	EE943	EE610	EE611	EE612	EE613
EE614	EE615	EE616	EE617	EE618	EE619	EE620	EE621	EE622	EE638	EE639	EE640	EE641
EE642	EE643	EE644	EE645	EE634	EE635	EE636	EE637	EE520	EE628	EE629	EE630	EE631
EE632	EE633	EE623	EE624	EE625	EE626	EE627	EE649	EE650	EE651	EE652	EE653	EE654
EE655	EE656	EE657	EE658	EE659	EE660	EE661	EE662	EE663	EE664	EE665	EE666	EE667
EE668	EE669	EE670	EE403	EE404	EE405	EE407	EE408	EE409	EE410	EE411	EE412	EE413
EE414	EE415	EE583	EE584	EE387	EE388	EE386	EE392	EE393	EE394	EE398	EE399	EE874
EE875	EE876	EE416	EE417	EE418	EE419	EE420	EE421	EE422	EE423	EE424	EE425	EE426
EE877	EE878	EE886	EE433	EE887	EE888	EE889	EE890	EE891	EE892	EE893	EE894	EE895
EE896	EE960	EE897	EE898	EE884	EE885	EE427	EE1005	EE434	EE429	EE430	EE431	EE998
EE432	EE997	EE433	EE428	EE1006	EE900	EE1000	EE879	EE901	EE882	EE406	EE1001	EE880
EE881	EE1002	EE899	EE986	EE833	EE902	EE996	EE991	EE992	EE435	EE961	EE903	EE904
EE905	EE438	EE439	EE962	EE442	EE446	EE447	EE448	EE963	EE964	EE906	EE440	EE907
EE908	EE909	EE466	EE467	EE468	EE469	EE470	EE471	EE472	EE473	EE474	EE475	EE476
EE477	EE478	EE479	EE480	EE481	EE482	EE483	EE484	EE485	EE486	EE487	EE488	EE489
EE490	EE491	EE492	EE493	EE494	EE495	EE916	EE917	EE918	EE919	EE920	EE921	EE922
EE923	EE496	EE497	EE498	EE499	EE507	EE508	EE959	EE512	EE513	EE514	EE521	EE522
EE572	EE573	EE574	EE575	EE576	EE577	EE578	EE579	EE580	EE581	EE582	EE603	EE604
EE605	EE606	EE607	EE608	EE609	EE671	EE672	EE673	EE674	EE675	EE676	EE677	EE678
EE679	EE680	EE681	EE682	EE683	EE684	EE685	EE686	EE687	EE688	EE689	EE690	EE691
EE692	EE693	EE694	EE695	EE696								

Table D.5: All unique item IDs for the available items in Q6.

All Available Item IDs from Q6												
GG486	GG487	GG488	GG491	GG492	GG493	GG494	GG495	GG496	GG497	GG498	GG499	GG500
GG501	GG502	GG606	GG607	GG608	GG609	GG610	GG611	GG509	GG510	GG511	GG512	GG513
GG450	GG451	GG449	GG514	GG515	GG516	GG517	GG518	GG519	GG520	GG521	GG612	GG613
GG614	GG615	GG435	GG600	GG601	GG602	GG603	GG604	GG605	GG522	GG523	GG524	GG525
GG526	GG527	GG528	GG529	GG530	GG531	GG532	GG533	GG534	GG535	GG536	GG537	GG538
GG539	GG540	GG541	GG489	GG490	GG222	GG223	GG224	GG225	GG237	GG238	GG239	GG240
GG241	GG242	GG226	GG227	GG228	GG229	GG230	GG383	GG384	GG385	GG390	GG391	GG392
GG393	GG394	GG395	GG396	GG397	GG597	GG27	GG28	GG231	GG232	GG233	GG234	GG235
GG236	GG243	GG244	GG245	GG246	GG247	GG248	GG249	GG250	GG251	GG252	GG253	GG254
GG592	GG255	GG256	GG257	GG258	GG259	GG260	GG261	GG262	GG263	GG264	GG265	GG266
GG267	GG268	GG269	GG270	GG271	GG272	GG273	GG274	GG275	GG276	GG277	GG278	GG279
GG280	GG281	GG282	GG283	GG284	GG285	GG286	GG287	GG288	GG289	GG290	GG291	GG292
GG293	GG294	GG299	GG300	GG301	GG302	GG303	GG304	GG305	GG306	GG307	GG308	GG309
GG310	GG311	GG312	GG313	GG314	GG315	GG316	GG317	GG318	GG319	GG320	GG321	GG322
GG323	GG324	GG325	GG326	GG327	GG328	GG329	GG330	GG331	GG332	GG333	GG334	GG335
GG336	GG337	GG338	GG339	GG340	GG341	GG342	GG343	GG344	GG345	GG346	GG347	GG348
GG349	GG350	GG351	GG352	GG353	GG354	GG355	GG356	GG357	GG358	GG359	GG360	GG361
GG362	GG363	GG364	GG365	GG366	GG367	GG236	GG436	GG437	GG438	GG439	GG440	GG441
GG442	GG443	GG444	GG445	GG446	GG447	GG634	GG635	GG636	GG637	GG638	GG639	GG640
GG641	GG642	GG643	GG644	GG645	GG646	GG647	GG648	GG649	GG650	GG651	GG652	GG653
GG654	GG655	GG656	GG657	GG479	GG480	GG481	GG482	GG483	GG484	GG485	GG503	GG504
GG505	GG506	GG507	GG508	GG616	GG617	GG618	GG619	GG620	GG621	GG622	GG623	GG624

Appendix E

Model Specifications

Here we list all sets of hyperparameter combinations needed to reproduce our results for the different datasets constructed. We list the different combinations in the same order as they appear in chapter 9 to make it clear which experiment the hyperparameter combinations belong to.

E.1 Experiment 1: Prediction Errors

For the prediction errors presented in this section in the result chapter we conducted one hyperparameter search per algorithm per dataset for one time step, and applied it to the two remaining time points. A total of 12 searches were made. The remaining of this sections is split into five subsections, one for each dataset constructed. The sets of hyperparameters belonging to a model are presented in a table.

E.1.1 Dataset with Aggregated Features from a Subset Selection

Table E.1 holds the different sets of hyperparameters for the elastic net, neural network and XGBoost model. The hyperparameters were used when we made predictions on the dataset containing aggregated features.

Table E.1: All hyperparameters and their values used in making the predictions in Table 9.1 in chapter 9 for the elastic net, neural network and XGBoost. The values were used in all time points.

Supervised Learning Algorithm					
Elastic Net		Neural Net		XGBoost	
Hyperparamter	Value	Hyperparamter	Value	Hyperparamter	Value
`alpha`	0.0071365	Optimizer	Adam	`subsample`	0.3
`l1_ratio`	0.0955441	Learning rate	0.001	`n_estimators`	10000
		Hidden layers	3	`min_child_weight`	3
		Number of Nodes	[500,500,100]	`max_depth`	6
		Weight Decay	False	`colsample_bytree`	0.4
		Drop-out	False	`eta`	0.01
		Epochs	1000	`early_stopping_rounds`	2500
		Batch size	64		
		Activation	ReLU		

E.1.2 Dataset Consisting of the Single Items from Aggregated Features

The hyperparameter values used for predicting the mean SCL when the data was made up by the single items from the aggregated features are in Table E.2.

Table E.2: All hyperparameters and their values used in making the predictions in Table 9.2 in chapter 9 for the elastic net, neural network and XGBoost. The values were used in all time points.

Supervised Learning Algorithm					
Elastic Net		Neural Net		XGBoost	
Hyperparamter	Value	Hyperparamter	Value	Hyperparamter	Value
`alpha`	0.3974877	Optimizer	SGD	`subsample`	0.5
`l1_ratio`	0.0292939	Learning rate	0.001	`n_estimators`	20000
		Hidden layers	6	`min_child_weight`	3
		Number of Nodes	[1000,1000,1000,100,1000,10]	`max_depth`	7
		Weight Decay	False	`colsample_bytree`	0.2
		Drop-out	False	`eta`	0.005
		Epochs	100	`early_stopping_rounds`	5000
		Batch size	32		
		Activation	tanh		

E.1.3 Dataset Constituted of Principal Components

All sets of hyperparameters values used when predicting the mean SCL value at all time points with the principal components that made up 95% of the explained variance are given in Table E.3.

Table E.3: All hyperparameters and their values used in making the predictions in Table 9.3 in chapter 9 for the elastic net, neural network and XGBoost. The values were used in all time points.

Supervised Learning Algorithm					
Elastic Net		Neural Net		XGBoost	
Hyperparamter	Value	Hyperparamter	Value	Hyperparamter	Value
`alpha`	0.0344372	Optimizer	Adam	`subsample`	0.4
`l1_ratio`	0.0449546	Learning rate	0.001	`n_estimators`	5000
		Hidden layers	6	`min_child_weight`	4
		Number of Nodes	[10,500,100,100,10,100]	`max_depth`	6
		Weight Decay	True	`colsample_bytree`	0.3
		Drop-out	False	`eta`	0.005
		Epochs	500	`early_stopping_rounds`	1250
		Batch size	32		
		Activation	relu		

E.1.4 Dataset with All Available Items

All sets of hyperparameters values used when predicting the mean SCL value at all time points with all available features are given in Table E.4.

Table E.4: All hyperparameters and their values used in making the predictions in Table 9.4 in chapter 9 for the elastic net, neural network and XGBoost. The values were used in all time points.

Supervised Learning Algorithm					
Elastic Net		Neural Net		XGBoost	
Hyperparamter	Value	Hyperparamter	Value	Hyperparamter	Value
`alpha`	0.0096337	Optimizer	SGD	`subsample`	0.5
`l1_ratio`	0.3867060	Learning rate	0.01	`n_estimators`	1000
		Hidden layers	6	`min_child_weight`	5
		Number of Nodes	[1000,1000,10,1000,500,10]	`max_depth`	7
		Weight Decay	True	`colsample_bytree`	0.3
		Drop-out	False	`eta`	0.01
		Epochs	100	`early_stopping_rounds`	250
		Batch size	32		
		Activation	relu		

E.1.5 Dataset with Correlated Items

All sets of hyperparameters values used when predicting the mean SCL value at all time points with the features that had an absolute correlation of 0.35 or higher with the target variable are shown in Table E.5.

Table E.5: All hyperparameters and their values used in making the predictions in Table 9.4 in chapter 9 for the elastic net, neural network and XGBoost. The values were used in all time points.

Supervised Learning Algorithm					
Elastic Net		Neural Net		XGBoost	
Hyperparamter	Value	Hyperparamter	Value	Hyperparamter	Value
`alpha`	0.0073505	Optimizer	SGD	`subsample`	0.5
`l1_ratio`	0.630267	Learning rate	0.001	`n_estimators`	10000
		Hidden layers	6	`min_child_weight`	5
		Number of Nodes	[1000,1000,1000,100,1000,10]	`max_depth`	7
		Weight Decay	False	`colsample_bytree`	0.2
		Drop-out	False	`eta`	0.005
		Epochs	100	`early_stopping_rounds`	250
		Batch size	64		
		Activation	tanh		

E.2 Experiment 2: Prediction Errors

When we ranked groups of features across the models for each specific time point we did a separate hyperparameter grid search for each time. Table E.6 list the combinations found for

predicting the mean SCL score at Q4, while Table E.7 and E.8 exhibits the hyperparameter values for Q5 and Q6 respectively.

Table E.6: All hyperparameters and their values used in making the predictions in Table 9.9 in chapter 9 for the elastic net, neural network and XGBoost at Q4.

Supervised Learning Algorithm					
Elastic Net		Neural Net		XGBoost	
Hyperparamter	Value	Hyperparamter	Value	Hyperparamter	Value
`alpha`	0.0071364	Optimizer	SGD	`subsample`	0.5
`l1_ratio`	0.0955411	Learning rate	0.001	`n_estimators`	20000
		Hidden layers	6	`min_child_weight`	1
		Number of Nodes	[100,500,500,500,500,10000]	`max_depth`	5
		Weight Decay	True	`colsample_bytree`	0.2
		Drop-out	False	`eta`	0.005
		Epochs	100	`early_stopping_rounds`	5000
		Batch size	64		
		Activation	relu		

Table E.7: All hyperparameters and their values used in making the predictions in Table 9.9 in chapter 9 for the elastic net, neural network and XGBoost at Q5.

Supervised Learning Algorithm					
Elastic Net		Neural Net		XGBoost	
Hyperparamter	Value	Hyperparamter	Value	Hyperparamter	Value
`alpha`	0.0071364	Optimizer	SGD	`subsample`	0.5
`l1_ratio`	0.0955411	Learning rate	0.001	`n_estimators`	10000
		Hidden layers	6	`min_child_weight`	5
		Number of Nodes	[100,500,500,500,500,10000]	`max_depth`	6
		Weight Decay	True	`colsample_bytree`	0.5
		Drop-out	False	`eta`	0.005
		Epochs	100	`early_stopping_rounds`	2500
		Batch size	64		
		Activation	relu		

From the three tables above it becomes evident that the grid search for the elastic nets and neural networks was performed using the same seed, while this was not the case for the XGBoost model. By changing the seed the different combinations could possibly yield better results.

Table E.8: All hyperparameters and their values used in making the predictions in Table 9.9 in chapter 9 for the elastic net, neural network and XGBoost at Q6.

Supervised Learning Algorithm					
Elastic Net		Neural Net		XGBoost	
Hyperparamter	Value	Hyperparamter	Value	Hyperparamter	Value
`alpha`	0.0071364	Optimizer	SGD	`subsample`	0.3
`l1_ratio`	0.0955411	Learning rate	0.001	`n_estimators`	2000
		Hidden layers	6	`min_child_weight`	5
		Number of Nodes	[100,500,500,500,500,10000]	`max_depth`	7
		Weight Decay	True	`colsample_bytree`	0.1
		Drop-out	False	`eta`	0.01
		Epochs	100	`early_stopping_rounds`	500
		Batch size	64		
		Activation	relu		

Bibliography

[1] Brian D Doss et al. "The effect of the transition to parenthood on relationship quality: an 8-year prospective study." eng. In: *Journal of personality and social psychology* 96.3 (Mar. 2009), pp. 601–619. ISSN: 0022-3514 (Print). DOI: 10.1037/a0013969.

[2] Gunvor Marie Dyrdal et al. "Can a Happy Relationship Predict a Happy Life? A Population-Based Study of Maternal Well-Being During the Life Transition of Pregnancy, Infancy, and Toddlerhood." In: *Journal of Happiness Studies* 12.6 (2011), pp. 947–962. ISSN: 13894978. DOI: 10.1007/s10902-010-9238-2.

[3] Sara McLanahan and Julia Adams. "Parenthood and psychological well-being." In: *Annual Review of Sociology* 13 (1987), pp. 237–257. ISSN: 1545-2115(Electronic),0360-0572(Print). DOI: 10.1146/annurev.so.13.080187.001321.

[4] Lesley Leeds and Isabel Hargreaves. "The psychological consequences of childbirth." In: *Journal of Reproductive and Infant Psychology* 26.2 (2008), pp. 108–122. ISSN: 02646838. DOI: 10.1080/02646830701688299.

[5] Michael W. O'Hara and Annette M. Swain. "Rates and risk of postpartum depression - A meta-analysis." In: *International Review of Psychiatry* 8.1 (1996), pp. 37–54. ISSN: 09540261. DOI: 10.3109/09540269609037816.

[6] Samantha Meltzer-Brody. "New insights into perinatal depression: Pathogenesis and treatment during pregnancy and postpartum." In: *Dialogues in Clinical Neuroscience* 13.1 (2011), pp. 89–100. ISSN: 12948322. DOI: 10.31887/dcns.2011.13.1/smbrody.

[7] Michael O'Hara. "Postpartum Depression: What We Know." In: *Journal of Clinical Psychology* 65 (2009), pp. 1258–1269. DOI: 10.1002/jclp.20644. URL: https://onlinelibrary.wiley.com/doi/epdf/10.1002/jclp.20644.

[8] Abdel Hady El-Gilany, Ghada O. Elkhawaga, and Bernadet B. Sarraf. "Depression and its associated factors among elderly: A community-based study in Egypt." In: *Archives of Gerontology and Geriatrics* 77.April (2018), pp. 103–107. ISSN: 18726976. DOI: 10.1016/j.archger.2018.04.011. URL: https://doi.org/10.1016/j.archger.2018.04.011.

[9] Nichole Fairbrother et al. "Perinatal anxiety disorder prevalence and incidence." eng. In: *Journal of affective disorders* 200 (Aug. 2016), pp. 148–155. ISSN: 1573-2517 (Electronic). DOI: 10.1016/j.jad.2015.12.082.

[10] Sandra Nakić Radoš, Meri Tadinac, and Radoslav Herman. "Anxiety During Pregnancy and Postpartum: Course, Predictors and Comorbidity with Postpartum Depression." eng. In: *Acta clinica Croatica* 57.1 (Mar. 2018), pp. 39–51. ISSN: 0353-9466. DOI: 10.20471/acc.2017.56.04.05. URL: https://pubmed.ncbi.nlm.nih.gov/30256010https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6400346/.

[11] Michelle P. Zappas, Kathleen Becker, and Benita Walton-Moss. "Postpartum Anxiety." In: *Journal for Nurse Practitioners* 17.1 (2021), pp. 60–64. ISSN: 15554155. DOI: 10.1016/j.nurpra.2020.08.017. URL: https://doi.org/10.1016/j.nurpra.2020.08.017.

[12] Corinna Reck et al. “Effects of postpartum anxiety disorders and depression on maternal self-confidence.” eng. In: *Infant behavior & development* 35.2 (Apr. 2012), pp. 264–272. ISSN: 1934-8800 (Electronic). DOI: 10.1016/j.infbeh.2011.12.005.

[13] Roberta A Mancuso et al. “Maternal prenatal anxiety and corticotropin-releasing hormone associated with timing of delivery.” eng. In: *Psychosomatic medicine* 66.5 (2004), pp. 762–769. ISSN: 1534-7796 (Electronic). DOI: 10.1097/01.psy.0000138284.70670.d5.

[14] V. Lindahl, J. L. Pearson, and L. Colpe. “Prevalence of suicidality during pregnancy and the postpartum.” In: *Archives of Women’s Mental Health* 8.2 (2005), pp. 77–87. ISSN: 14341816. DOI: 10.1007/s00737-005-0080-1.

[15] Niloufer Sultan Ali et al. “Impact of postpartum anxiety and depression on child’s mental development from two peri-urban communities of Karachi, Pakistan: a quasi-experimental study.” eng. In: *BMC psychiatry* 13 (Oct. 2013), p. 274. ISSN: 1471-244X (Electronic). DOI: 10.1186/1471-244X-13-274.

[16] Lisa A. Serbin et al. “The influence of parenting on early childhood health and health care utilization.” In: *Journal of Pediatric Psychology* 39.10 (2014), pp. 1161–1174. ISSN: 1465735X. DOI: 10.1093/jpepsy/jsu050.

[17] Ida Kathrine Gravensteen et al. “Anxiety, depression and relationship satisfaction in the pregnancy following stillbirth and after the birth of a live-born baby: A prospective study.” In: *BMC Pregnancy and Childbirth* 18.1 (2018), pp. 1–10. ISSN: 14712393. DOI: 10.1186/s12884-018-1666-8.

[18] Karine Eid et al. “Perinatal Depression and Anxiety in Women With Multiple Sclerosis: A Population-Based Cohort Study.” In: *Neurology* 96.23 (2021), e2789–e2800. ISSN: 1526632X. DOI: 10.1212/WNL.0000000000012062.

[19] Marte Helene Bjørk et al. “Depression and anxiety in women with epilepsy during pregnancy and after delivery: A prospective population-based cohort study on frequency, risk factors, medication, and prognosis.” In: *Epilepsia* 56.1 (2015), pp. 28–39. ISSN: 15281167. DOI: 10.1111/epi.12884.

[20] Marie F. Sørbø et al. “Adult physical, sexual, and emotional abuse and postpartum depression, a population based, prospective study of 53,065 women in the norwegian mother and child cohort study.” In: *BMC Pregnancy and Childbirth* 14.1 (2014), pp. 1–9. ISSN: 14712393. DOI: 10.1186/1471-2393-14-316.

[21] Carl Van Walraven and Robert G. Hart. “Leave ’em alone - Why continuous variables should be analyzed as such.” In: *Neuroepidemiology* 30.3 (2008), pp. 138–139. ISSN: 02515350. DOI: 10.1159/000126908.

[22] Che Wan Jasimah Bt Wan Mohamed Radzi, Hashem Salarzadeh Jenatabadi, and Nadia Samsudin. “Postpartum depression symptoms in survey-based research: a structural equation analysis.” In: *BMC Public Health* 21.1 (2021), pp. 1–12. ISSN: 14712458. DOI: 10.1186/s12889-020-09999-2.

[23] Laura Miller. “Postpartum depression.” In: *Journal of Nursing* 60.6 (2002), pp. 22–26. ISSN: 0047262X. DOI: 10.6224/JN.60.6.22.

[24] Gwen Stern and Laurence Kruckman. “Multi-disciplinary perspectives on post-partum depression: An anthropological critique.” In: *Social S* 17.15 (1989), pp. 1027–1041. URL: https://doi.org/10.1016/0277-9536(83)90408-2.

[25] Open Science Collaboration. “Estimating the reproducibility of psychological science.” In: *Science* 349 (2015), aac4716. DOI: 10.1126/science.aac4716. URL: https://www.science.org/doi/abs/10.1126/science.aac4716.

[26] Graziella Orrù et al. "Machine learning in psychometrics and psychological research." In: *Frontiers in Psychology* 10.January (2020), pp. 1–10. ISSN: 16641078. DOI: 10.3389/fpsyg.2019.02970.

[27] Tal Yarkoni and Jacob Westfall. "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning." In: *Perspectives on Psychological Science* 12.6 (2017), pp. 1100–1122. ISSN: 17456924. DOI: 10.1177/1745691617693393.

[28] María J. Blanca, Rafael Alarcón, and Roser Bono. "Current practices in data analysis procedures in psychology: What has changed?" In: *Frontiers in Psychology* 9.DEC (2018), pp. 1–12. ISSN: 16641078. DOI: 10.3389/fpsyg.2018.02558.

[29] Jacob M. Marszalek et al. "Sample size in psychological research over the past 30 years." In: *Perceptual and Motor Skills* 112.2 (2011), pp. 331–348. ISSN: 00315125. DOI: 10.2466/03.11.PMS.112.2.331-348.

[30] Kai Sassenberg and Lara Ditrich. "Research in Social Psychology Changed Between 2011 and 2016: Larger Sample Sizes, More Self-Report Measures, and More Online Studies." In: *Advances in Methods and Practices in Psychological Science* 2.2 (2019), pp. 107–114. ISSN: 2515-2459. DOI: 10.1177/2515245919838781.

[31] Kevin Grimm, Ross Jacobucci, and John McArdle. *Big data methods and psychological science*. 2017. URL: https://www.apa.org/science/about/psa/2017/01/big-data-methods.

[32] Robert M. Kaplan, David A. Chambers, and Russell E. Glasgow. "Big data and large sample size: A cautionary note on the potential for bias." In: *Clinical and Translational Science* 7.4 (2014), pp. 342–346. ISSN: 17528062. DOI: 10.1111/cts.12178.

[33] Per Magnus et al. "Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa)." In: *International Journal of Epidemiology* 45.2 (2016), pp. 382–388. ISSN: 14643685. DOI: 10.1093/ije/dyw029.

[34] Ragnhild Eek Brandlistuen et al. "Prenatal paracetamol exposure and child neurodevelopment: A sibling-controlled cohort study." In: *International Journal of Epidemiology* 42.6 (2013), pp. 1702–1713. ISSN: 03005771. DOI: 10.1093/ije/dyt183.

[35] Roger Ekeberg Henriksen, Torbjørn Torsheim, and Frode Thuen. "Relationship satisfaction reduces the risk of maternal infectious diseases in pregnancy: The Norwegian mother and child cohort study." In: *PLoS ONE* 10.1 (2015), pp. 1–10. ISSN: 19326203. DOI: 10.1371/journal.pone.0116796.

[36] Maria Magnus et al. "Grandmother's smoking when pregnant with the mother and asthma in the grandchild: the Norwegian Mother and Child Cohort Study." In: *Thorax* 70.3 (2015), pp. 237–243. DOI: 10.1136/thoraxjnl-2014-206438. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5034931/pdf/nihms816826.pdf.

[37] *Publications from the Norwegian Mother, Father and Child Cohort Study*. 2021. URL: https://www.fhi.no/en/studies/moba/for-forskere-artikler/publications/.

[38] Xiaoyu Che et al. "Maternal mid-gestational and child cord blood immune signatures are strongly associated with offspring risk of ASD." In: *Molecular Psychiatry* January (2022). ISSN: 14765578. DOI: 10.1038/s41380-021-01415-4.

[39] Mohiuddeen Khan and Kanishk Srivastava. "Regression model for better generalization and regression analysis." In: *ACM International Conference Proceeding Series* January 2020 (2020), pp. 30–33. DOI: 10.1145/3380688.3380691.

[40] Danilo Bzdok, Naomi Altman, and Martin Krzywinski. "Points of Significance: Statistics versus machine learning." In: *Nature Methods* 15.4 (2018), pp. 233–234. ISSN: 15487105. DOI: 10.1038/nmeth.4642. URL: http://dx.doi.org/10.1038/nmeth.4642.

[41] Mattea Romano et al. “Postpartum period: three distinct but continuous phases.” eng. In: *Journal of prenatal medicine* 4.2 (Apr. 2010), pp. 22–25. ISSN: 1971-3290 (Electronic).

[42] *What is the Norwegian Mother, Father and Child Cohort Study?* 2021. URL: https://www.fhi.no/en/studies/moba/what-is-the-norwegian-mother-and-child-cohort-study/.

[43] Zahra M Clayborne et al. “Prenatal work stress is associated with prenatal and postnatal depression and anxiety: Findings from the Norwegian Mother, Father and Child Cohort Study (MoBa).” eng. In: *Journal of affective disorders* 298.Pt A (Feb. 2022), pp. 548–554. ISSN: 1573-2517 (Electronic). DOI: 10.1016/j.jad.2021.11.024.

[44] Espen Moen Eilertsen et al. “Parental Prenatal Symptoms of Depression and Offspring Symptoms of ADHD: A Genetically Informed Intergenerational Study.” In: *Journal of Attention Disorders* 25.11 (2021), pp. 1554–1563. ISSN: 15571246. DOI: 10.1177/1087054720914386.

[45] Wonuola A. Akingbuwa et al. “Genetic Associations between Childhood Psychopathology and Adult Depression and Associated Traits in 42998 Individuals: A Meta-analysis.” In: *JAMA Psychiatry* 77.7 (2020), pp. 715–728. ISSN: 2168622X. DOI: 10.1001/jamapsychiatry.2020.0527.

[46] Eivind Ystrom et al. “Maternal Symptoms of Anxiety and Depression and Child Nocturnal Awakenings at 6 and 18 Months.” In: *Journal of Pediatric Psychology* 42.10 (2017), pp. 1156–1164. ISSN: 1465735X. DOI: 10.1093/jpepsy/jsx066.

[47] Seung Yong Han, Alexandra A. Brewis, and Amber Wutich. “Body image mediates the depressive effects of weight gain in new mothers, particularly for women already obese: Evidence from the Norwegian Mother and Child Cohort Study.” In: *BMC Public Health* 16.1 (2016), pp. 1–10. ISSN: 14712458. DOI: 10.1186/s12889-016-3363-8. URL: http://dx.doi.org/10.1186/s12889-016-3363-8.

[48] *International Statistical Classification of Diseases and Related Health Problems*. 11th ed. World Health Organization, 2019. URL: https://icd.who.int/en.

[49] *Depression*. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

[50] A J Baxter et al. “Global prevalence of anxiety disorders: a systematic review and meta-regression.” eng. In: *Psychological medicine* 43.5 (May 2013), pp. 897–910. ISSN: 1469-8978 (Electronic). DOI: 10.1017/S003329171200147X.

[51] Priyamadhaba Behera et al. “Screening instruments for assessment of depression.” In: *Indian Journal of Medical Specialities* 8.1 (2017), pp. 31–37. ISSN: 09762884. DOI: 10.1016/j.injms.2016.11.003. URL: http://dx.doi.org/10.1016/j.injms.2016.11.003.

[52] Rudolf Uher et al. “Self-report and clinician-rated measures of depression severity: Can one replace the other?” In: *Depression and Anxiety* 29.12 (2012), pp. 1043–1049. ISSN: 10914269. DOI: 10.1002/da.21993.

[53] Brett D. Thombs et al. “Addressing overestimation of the prevalence of depression based on self-report screening questionnaires.” In: *Cmaj* 190.2 (2018), E44–E49. ISSN: 14882329. DOI: 10.1503/cmaj.170691.

[54] Leonard R Derogatis et al. *The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory*. US, 1974. DOI: 10.1002/bs.3830190102.

[55] R Likert. “A technique for the measurement of attitudes.” In: *Archives of Psychology* 22 140 (1932), p. 55.

[56] Tanguy Le Bris. “The Hopkins symptoms checklist in 25 items : translations in Castilian, Galician, Catalan, French, Greek, Italian, Polish, Bulgarian and Croatian synthesis. Life.” In: *Life Science* q-bio (2017), dumas–01537933.

[57] Bjørn Heine Strand et al. "Measuring the mental health status of the Norwegian population: A comparison of the instruments SCL-25, SCL-10, SCL-5 and MHI-5 (SF-36)." In: *Nordic Journal of Psychiatry* 57.2 (2003), pp. 113–118. ISSN: 08039488. DOI: 10.1080/08039480310000932.

[58] *Questions Documentation Questionnaire 4 6 months old.* 2016. URL: https://www.fhi.no/globalassets/dokumenterfiler/studier/den-norske-mor-far-og-barn--undersokelsenmoba/instrumentdokumentasjon/instrument-documentation-q4.pdf.

[59] Louise Emanuel. "The effects of post-natal depression on the child." In: *Psycho-analytic Psychotherapy in South Africa* 7.1 (1999), pp. 50–67. URL: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc3&NEWS=N&AN=2000-13252-003.

[60] Janice H. Goodman. "Paternal postpartum depression, its relationship to maternal postpartum depression, and implications for family health." In: *Journal of Advanced Nursing* 45.1 (2004), pp. 26–35. ISSN: 03092402. DOI: 10.1046/j.1365-2648.2003.02857.x.

[61] Michael W. O'Hara et al. "Controlled prospective study of postpartum mood disorders: Comparison of childbearing and nonchildbearing women." In: *Journal of Abnormal Psychology* 99.1 (1990), pp. 3–15. ISSN: 0021-843X. DOI: 10.1037//0021-843x.99.1.3.

[62] P. J. Cooper et al. "Non-psychotic psychiatric disorder after childbirth. A prospective study of prevalence, incidence, course and nature." In: *British Journal of Psychiatry* 152.JUN. (1988), pp. 799–806. ISSN: 00071250. DOI: 10.1192/bjp.152.6.799.

[63] Palo Almond. "Postnatal depression: A global public health perspective." In: *Perspectives in Public Health* 129.5 (2009), pp. 221–227. ISSN: 17579139. DOI: 10.1177/1757913909343882.

[64] Kobra Falah-Hassani et al. "Prevalence of postpartum depression among immigrant women: A systematic review and meta-analysis." In: *Journal of Psychiatric Research* 70 (2015), pp. 67–82. ISSN: 18791379. DOI: 10.1016/j.jpsychires.2015.08.010. URL: http://dx.doi.org/10.1016/j.jpsychires.2015.08.010.

[65] Maryam Ghaedrahmati et al. "Postpartum depression risk factors: A narrative review." In: *Journal of Education and Health Promotion* 6.60 (2017). DOI: 10.4103/jehp.jehp{_}9{_}16.

[66] A F Bell et al. "Childbirth and symptoms of postpartum depression and anxiety: a prospective birth cohort study." eng. In: *Archives of women's mental health* 19.2 (Apr. 2016), pp. 219–227. ISSN: 1435-1102 (Electronic). DOI: 10.1007/s00737-015-0555-7.

[67] Marc Gellman and Rick Turner J. *Encyclopedia of Behavioral Medicine.* 2013. ISBN: 9781441910059. DOI: 10.1007/978-1-4419-1005-9.

[68] David Watson and Lee A. Clark. "Negative affectivity: The disposition to experience aversive emotional states." In: *Psychological Bulletin* 96.3 (1984), pp. 465–490. ISSN: 00332909. DOI: 10.1037/0033-2909.96.3.465.

[69] Albert Bandura. "Self-efficacy: Toward a Unifying Theory of Behavioral Change." In: *Psychological Review* 84.2 (1977), pp. 191–215.

[70] Isabelle Guyon and André Elisseeff. "An Introduction to Variable and Feature Selection." In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[71] Tanya P Garcia and Karen Marder. "Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model." In: *Current Neurology and Neuroscience Reports* 17.2 (2017). ISSN: 15346293. DOI: 10.1007/s11910-017-0723-4.

[72] Steven F. Sawyer. "Analysis of Variance: The Fundamental Concepts." In: *Journal of Manual \& Manipulative Therapy* 17.2 (2009), 27E–38E. ISSN: 1066-9817. DOI: 10.1179/jmt.2009.17.2.27e.

[73] Ita Kreft and Jan de Leeuw. *Introducing Multilevel Modeling*. London, 1998. DOI: 10.4135/9781849209366. URL: https://methods.sagepub.com/book/introducing-multilevel-modeling.

[74] Moritz Herle et al. "Identifying typical trajectories in longitudinal data: modelling strategies and interpretations." In: *European Journal of Epidemiology* 35.3 (2020), pp. 205–222. ISSN: 15737284. DOI: 10.1007/s10654-020-00615-6. URL: https://doi.org/10.1007/s10654-020-00615-6.

[75] Sara Boslaugh. "Structural Equation Modeling." In: (2008), pp. 1005–1008. DOI: http://dx.doi.org/10.4135/9781412953948.n443.

[76] Daniel McNeish and Tyler Matta. "Differentiating between mixed-effects and latent-curve approaches to growth modeling." In: *Behavior Research Methods* 50.4 (2018), pp. 1398–1414. ISSN: 15543528. DOI: 10.3758/s13428-017-0976-5.

[77] Emily A. Blood et al. "Performance of mixed effects models in the analysis of mediated longitudinal data." In: *BMC Medical Research Methodology* 10 (2010). ISSN: 14712288. DOI: 10.1186/1471-2288-10-16.

[78] *Health Informatics*. 2021. URL: https://en.wikipedia.org/w/index.php?title=Health_informatics&oldid=1048230692.

[79] Daniele Ravi et al. "Deep Learning for Health Informatics." In: *IEEE Journal of Biomedical and Health Informatics* 21.1 (2017), pp. 4–21. ISSN: 21682194. DOI: 10.1109/JBHI.2016.2636665.

[80] David Bellamy, Leo Celi, and Andrew L. Beam. "Evaluating Progress on Machine Learning for Longitudinal Electronic Healthcare Data." In: (2020). URL: http://arxiv.org/abs/2010.01149.

[81] Benjamin Shickel et al. "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis." In: *IEEE Journal of Biomedical and Health Informatics* 22.5 (2018), pp. 1589–1604. ISSN: 21682194. DOI: 10.1109/JBHI.2017.2767063.

[82] Xuhong Li et al. "Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond." In: *arXiv* (2021). URL: https://arxiv.org/abs/2103.10689.

[83] Julia Amann et al. "Explainability for artificial intelligence in healthcare : a multidisciplinary perspective." In: *BMC Medical Informatics and Decision Making* (2020), pp. 1–9. ISSN: 1472-6947. DOI: 10.1186/s12911-020-01332-6. URL: https://doi.org/10.1186/s12911-020-01332-6.

[84] Pantelis Linardatos and Vasilis Papastefanopoulos. "Explainable AI : A Review of Machine Learning Interpretability Methods." In: *Entropy* 23.1 (2021). DOI: https://doi.org/10.3390/e23010018.

[85] Derek Doran, Sarah Schulz, and Tarek R Besold. "What Does Explainable AI Really Mean ? A New Conceptualization of Perspectives." In: *arXiv* (2017).

[86] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"." In: *AI Magazine* 38.3 (June 2017). DOI: 10.1609/aimag.v38i3.2741. URL: https://arxiv.org/abs/1606.08813.

[87] Leo Breiman. "Statistical modeling: The two cultures." In: *Statistical Science* 16.3 (2001), pp. 199–215. ISSN: 08834237. DOI: 10.1214/ss/1009213726.

[88] Leo Breiman et al. *Classification and Regression Trees*. Chapman and Hall, 1984, p. 368. ISBN: 9780412048418.

[89] Galit Shmueli. "To explain or to predict?" In: *Statistical Science* 25.3 (2010), pp. 289–310. ISSN: 08834237. DOI: 10.1214/10-STS330.

[90] Scott E. Maxwell, Michael Y. Lau, and George S. Howard. "Is psychology suffering from a replication crisis?: What does 'failure to replicate' really mean?" In: *American Psychologist* 70.6 (2015), pp. 487–498. ISSN: 0003066X. DOI: 10.1037/a0039400.

[91] John M. Aiken and H. J. Lewandowski. "Data sharing model for physics education research using the 70 000 response Colorado Learning Attitudes about Science Survey for Experimental Physics dataset." In: *Physical Review Physics Education Research* 17.2 (2021), p. 20144. ISSN: 2469-9896. DOI: 10.1103/physrevphyseducres.17.020144. URL: https://doi.org/10.1103/PhysRevPhysEducRes.17.020144.

[92] Jake M Hofman, Amit Sharma, and Duncan J Watts. "Prediction and explanation in social systems." In: 488.February (2017), pp. 486–488.

[93] Russell A. Poldrack, Grace Huckins, and Gael Varoquaux. "Establishment of Best Practices for Evidence for Prediction: A Review." In: *JAMA Psychiatry* 77.5 (2020), pp. 534–540. ISSN: 2168622X. DOI: 10.1001/jamapsychiatry.2019.3671.

[94] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." In: *Psychological Science* 22.11 (2011), pp. 1359–1366. ISSN: 14679280. DOI: 10.1177/0956797611417632.

[95] Wojciech Swiatkowski and BenoÎt Dompnier. "Replicability crisis in social psychology: Looking at the past to find new pathways for the future." In: *International Review of Social Psychology* 30.1 (2017), pp. 111–124. ISSN: 23978570. DOI: 10.5334/irsp.66.

[96] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*. 1st. Belmont, Massachusetts: Athena Scientific, 2002, p. 250. ISBN: 1-886529-40-X.

[97] S.S. Stevens. "On the Theory of Scales of Measurement." In: *Science* 103.2684 (1946), pp. 677–680. ISSN: 0036-8075.

[98] Aurélien G{\'e}ron. *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. O'Reilly, 2017, p. 119. ISBN: 9781491962299.

[99] Sergios Theodoridis and Konstantinos Koutroumbas. *An Introduction to Pattern Recognition: A MATLAB Approach*. Oxford: Academic Press, 2010, p. 215. ISBN: 9780123744869.

[100] Donald B. Rubin. "Inference and Missing Data." In: *Biometrika* 63.3 (1976), pp. 581–592. ISSN: 1471-2105.

[101] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007, p. 607. ISBN: 978-0-511-26878-6.

[102] Roderick Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. 3rd. Hoboken: Wiley, 2020, p. 449. ISBN: 9781118595695.

[103] Zhenhua Wang et al. "Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison." In: (2021). URL: http://arxiv.org/abs/2103.09316.

[104] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. 1st. Hoboken, 1987, p. 258. ISBN: 0-471-08705-X.

[105] Stef van Buuren. "Multiple imputation of discrete and continuous data by fully conditional specification." In: *Statistical Methods in Medical Research* 16.3 (2007), pp. 219–242. ISSN: 09622802. DOI: 10.1177/0962280206074463.

[106] Melissa Azur et al. "Multiple imputation by chained equations: what is it and how does it work?" In: *International Journal of Methods in Psychiatric Research* 1 (2011), pp. 40–49. DOI: 10.1002/mpr.329.

[107] Olanrewaju Akande, Fan Li, and Jerome Reiter. "An Empirical Comparison of Multiple Imputation Methods for Categorical Data." In: *American Statistician* 71.2 (2017), pp. 162–170. ISSN: 15372731. DOI: 10.1080/00031305.2016.1277158.

[108] Pankaj Mehta et al. "A high-bias, low-variance introduction to Machine Learning for physicists." In: *Physics Reports* 810 (2018), p. 23. DOI: 10.1016/j.physrep.2019.03.001.

[109] T. Chai and R. 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. ISSN: 19919603. DOI: 10.5194/gmd-7-1247-2014.

[110] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization." In: *Journal of Machine Learning Research* 13 (2012), pp. 281–305. ISSN: 15324435.

[111] B Efron. "Boostrap methods: Another look at the Jackknife." In: *Annals of Statistics* 7.1 (1979), pp. 1–26.

[112] Diederik P. Kingma and Jimmy Lei Ba. "Adam: A method for stochastic optimization." In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015), pp. 1–15.

[113] Naum Shor. *Nondifferentiable Optimization and Polynomial Problems*. Kiev: Springer, 1998, p. 335. ISBN: 978-1-4757-6015-6. DOI: 10.1007/978-1-4757-6015-6.

[114] Stephen Boyd and Jaehyun Park. "Subgradient Methods." 2014. URL: https://web.stanford.edu/class/ee364b/lectures/subgrad_method_notes.pdf.

[115] Jay L Devore and Kenneth N Berk. *Modern Mathematical Statistics with Applications*. 2nd. Springer, 2011. ISBN: 9781461403906. DOI: 10.1007/978-1-4614-0391-3. URL: http://www.springer.com/series/417.

[116] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." In: *Technometrics* 12.1 (1970), pp. 55–67. ISSN: 15372723. DOI: 10.1080/00401706.1970.10488634.

[117] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 2nd. Vol. 103. 2. Springer, 2017, p. 251. ISBN: 9780387848570. DOI: 10.2307/2980421.

[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* 67.5 (2005), pp. 301–320. ISSN: 13697412. DOI: 10.1111/j.1467-9868.2005.00527.x.

[119] Michael Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. URL: http://neuralnetworksanddeeplearning.com.

[120] David Rumelhart, Geoffrey Hinton, and Ronald Williams. "Learning Representations by Back-Propagating Errors." In: *Cognitive Modeling* 2 (1986), pp. 3–6. DOI: 10.7551/mitpress/1888.003.0013.

[121] Lu Lu et al. "Dying ReLU and initialization: Theory and numerical examples." In: *Communications in Computational Physics* 28.5 (2020), pp. 1671–1706. ISSN: 19917120. DOI: 10.4208/CICP.OA-2020-0165.

[122] Lutz Prechelt. “Early stopping - But when?” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7700 LECTU (2012), pp. 53–67. ISSN: 16113349. DOI: 10.1007/978-3-642-35289-8{_}5.

[123] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: deeplearningbook.org.

[124] Geoffrey E. Hinton et al. “Improving neural networks by preventing co-adaptation of feature detectors.” In: (2012), pp. 1–18. URL: http://arxiv.org/abs/1207.0580.

[125] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: https://jmlr.org/papers/volume15/srivastava14a.html.

[126] Thomas G. Dietterich. “Ensemble methods in machine learning.” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1857 LNCS (2000), pp. 1–15. ISSN: 16113349. DOI: 10.1007/3-540-45014-9{_}1.

[127] Tianqi Chen and Carlos Guestrin. “XGBoost: A scalable tree boosting system.” In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-Augu (2016), pp. 785–794. DOI: 10.1145/2939672.2939785.

[128] *XGBoost Parameters*. URL: https://xgboost.readthedocs.io/en/stable/parameter.html.

[129] Jerome H. Friedman. “Greedy function approximation: A gradient boosting machine.” In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232. ISSN: 00905364. DOI: 10.1214/aos/1013203451.

[130] Patricia Schreuder and Elin Alsaker. “The Norwegian Mother and Child Cohort Study (MoBa) – MoBa recruitment and logistics.” In: *Norsk Epidemiologi* 24 (2014), pp. 23–27. DOI: 10.5324/nje.v24i1-2.1754. URL: https://www.researchgate.net/publication/292518152_The_Norwegian_Mother_and_Child_Cohort_Study_MoBa_-_MoBa_recruitment_and_logistics.

[131] *Questionnaires from MoBa*. 2021. URL: https://www.fhi.no/en/studies/moba/for-forskere-artikler/questionnaires-from-moba/.

[132] *Questions Documentation Questionnaire 1 15 th week of gestation*. 2016. URL: https://www.fhi.no/globalassets/dokumenterfiler/studier/moba/dokumenter/instrument-documentation-q1.pdf.

[133] *Questions Documentation Questionnaire 3 30 th week of gestation*. 2016. URL: https://www.fhi.no/globalassets/dokumenterfiler/studier/den-norske-mor-far-og-barn--undersokelsenmoba/instrumentdokumentasjon/instrument-documentation-q3.pdf.

[134] K. S Kendler et al. “The lifetime history of major depression in women: reliability of diagnosis and heritability.” In: *Archives of General Psychiatry,* 50 (1993), pp. 863–870. DOI: 10.1001/archpsyc.1993.01820230054003..

[135] Ronald C Kessler et al. “Validity of the World Health Organization Adult ADHD Self-Report Scale (ASRS) Screener in a representative sample of health plan members.” eng. In: *International journal of methods in psychiatric research* 16.2 (2007), pp. 52–65. ISSN: 1049-8931 (Print). DOI: 10.1002/mpr.208.

[136] Ralf Schwarzer and Matthias Jerusalem. “Generalized Self-Efficacy scale.” In: *In J. Weinman, S. Wright, \& M. Johnston: Measures in health psychology: A user's portfolio. Causal and control beliefs* (1995), pp. 35–37.

[137] Izard Carroll et al. “Stability of emotion experiences and their relations to traits of personality.” In: *Journal of Personality and Social Psychology* 64.5 (1993), pp. 847–860.

[138] Morris Rosenberg. *Society and the Adolescent Self-Image.* Revised Ed. Wesleyan University Press., 1989.

[139] Ed Diener et al. "The Satisfaction With Life Scale." In: *Journal of Personality Assessment,* 41.1 (1985), pp. 71–75.

[140] The WHOQOL Group. "Development of the World Health Organization WHOQOL-BREF quality of life assessment." In: *Psychological Medicine* 28.3 (1998), pp. 551–558.

[141] JB Sanders et al. "Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II." In: *Addiction* 88 (1993), pp. 791–804.

[142] HR White and EW Labouvie. "Towards the assessment of adolescent problem drinking." In: *Journal of Studies on Alcohol* 50 (1989), pp. 30–37.

[143] Jane Squires, Diane D Bricker, and LaWanda Potter. *The ASQ user's guide / by Jane Squires, LaWanda Potter, and Diane Bricker.* 2nd ed. Baltimore, MD: Paul H. Brookes, 1999. ISBN: 155766367X.

[144] J. Bates, C. Freeland, and M. Lounsbury. "Measurement of infant difficultness. Infant characteristic questionnaire (ICQ)." In: *Child Development* 50.3 (1979), pp. 794–803.

[145] A. H Buss and R Plomin. *Temperament : Early developing personality traits.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1984. ISBN: 9781138823174.

[146] T Achenback. *Manual for the Child Behaviour Checklist/2-3 and 1992 Profile.* Burlington, VT : Dept. of Psychiatry, University of Vermont, 1992, p. 210. ISBN: 0938565206 9780938565208.

[147] C Diets et al. "Screening for autistic spectrum disorder in children aged 14-15 months. II: Population screening with the Early Screening of Autistic Traits Questionnaire (ESAT), design and general findings." In: *Journal of Autism and Developmental Disorders* 36 (2006), pp. 713–722.

[148] L Robins et al. "The Modified Checklist for Autism in Toddlers: An Initial Study Investigating the Early Detection of Autism and Pervasive Developmental Disorders." In: *Journal of Autism and Developmental Disorders* 161.6 (2001), pp. 131–144.

[149] *Old, Questions Documentation Questionnaire 5 18 months.* 2016. URL: https://www.fhi.no/globalassets/dokumenterfiler/studier/den-norske-mor-far-og-barn--undersokelsenmoba/instrumentdokumentasjon/instrument-documentation-q5.pdf.

[150] Synnve Schjolberg. *Test retest reliability of a screening checklist for Autism Spectrum disorders in young children.* Tech. rep. Boston, 2005.

[151] M Rutter, A Bailey, and C Lord. *SCQ The Social Communication Questionnaire: Manual.* Los Angeles: Western Psychological Services, 2003.

[152] R Goodman. "The Strengths and Difficulties Questionnaire: A Research Note." In: *Child Psychology and Psychiatry* 38 (1997), pp. 581–586.

[153] Farrokh Habibzadeh, Parham Habibzadeh, and Mahboobeh Yadollahie. "On determining the most appropriate test cut-off value: the case of tests with continuous results." eng. In: *Biochemia medica* 26.3 (Oct. 2016), pp. 297–307. ISSN: 1330-0962 (Print). DOI: 10.11613/BM.2016.034.

[154] Klaus Peter Ossenkopp and Dwight S. Mazmanian. "The principle of aggregation in psychobiological correlational research: An example from the open-field test." In: *Animal Learning \& Behavior* 13.4 (1985), pp. 339–344. ISSN: 00904996. DOI: 10.3758/BF03208007.

[155] Christopher J. Urban and Kathleen M. Gates. "Deep Learning: A Primer for Psychologists." In: *Psychological Methods* 26.6 (2021), pp. 743–773. ISSN: 1082989X. DOI: 10.1037/met0000374.

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

[157] Cort J. Willmott and Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." In: *Climate Research* 30.1 (2005), pp. 79–82. ISSN: 0936577X. DOI: 10.3354/cr030079.

[158] François Chollet. *Keras*. 2015. URL: https://keras.io.

[159] D.H Wolpert and W. G Macready. "No free lunch theorems for optimization." In: *IEEE Transactions on Evolutionary Computation* 1.1 (1997), pp. 67–82. DOI: 10.1109/4235.585893.

[160] Joanna J Arch. "Pregnancy-specific anxiety: which women are highest and what are the alcohol-related risks?" eng. In: *Comprehensive psychiatry* 54.3 (Apr. 2013), pp. 217–228. ISSN: 1532-8384 (Electronic). DOI: 10.1016/j.comppsych.2012.07.010.

[161] Sheila M. Marcus. "Depression during Pregnancy: Rates, Risks and Consequences." In: *Journal of Population Therapeutics \& Clinical Pharmacology* 16.1 (2009), pp. 15–22. URL: https://jptcp.com/index.php/jptcp/article/view/295.

[162] Statistisk Sentralbyrå. *Personer 16 år og over, etter høyeste fullførte utdanning, kjønn og alder. 1. oktober 2004*. Tech. rep. 2005, p. 1. URL: https://www.ssb.no/a/kortnavn/utniv/arkiv/tab-2005-08-26-03.html.

[163] Anne Enger et al. *Kjønn og lønn*. Tech. rep. 2008, p. 310. URL: http://www.regjeringen.no/nb/dep/bld/dok/nouer/2008/nou-2008-6/9/1.html?id=501163.

[164] Wensu Zhou et al. "Emotional problems in mothers of autistic children and their correlation with socioeconomic status and the children's core symptoms." eng. In: *Medicine* 98.32 (Aug. 2019), e16794. ISSN: 1536-5964 (Electronic). DOI: 10.1097/MD.0000000000016794.

[165] T D Gedeon. "Data mining of inputs: analysing magnitude and functional measures." eng. In: *International journal of neural systems* 8.2 (Apr. 1997), pp. 209–218. ISSN: 0129-0657 (Print). DOI: 10.1142/s0129065797000227.

[166] D. G Garson. "Interpreting neural-network connection weights." In: *Artificial Intelligence Expert* 6.4 (1991), pp. 46–51.

[167] Scott Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In: 2017. URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[168] Ravid Shwartz-Ziv and Amitai Armon. "Tabular data: Deep learning is not all you need." In: *Information Fusion* 81 (2022), pp. 84–90. ISSN: 15662535. DOI: 10.1016/j.inffus.2021.11.011.

[169] Jihoon Oh et al. "Identifying depression in the National Health and Nutrition Examination Survey data using a deep learning algorithm." In: *Journal of Affective Disorders* 257.July (2019), pp. 623–631. ISSN: 15732517. DOI: 10.1016/j.jad.2019.06.034. URL: https://doi.org/10.1016/j.jad.2019.06.034.

[170] O. Naggara et al. "Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms." In: *American Journal of Neuroradiology* 32.3 (2011), pp. 437–440. ISSN: 01956108. DOI: 10.3174/ajnr.A2425.

[171] Yuzhe Yang et al. “Delving into Deep Imbalanced Regression.” In: (2021). URL: http://arxiv.org/abs/2102.09554.

[172] T G Dietterich. “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms.” eng. In: *Neural computation* 10.7 (Sept. 1998), pp. 1895–1923. ISSN: 1530-888X (Electronic). DOI: 10.1162/089976698300017197.

[173] Fabio Sigrist. “Gaussian Process Boosting.” In: (2020), pp. 1–41. URL: http://arxiv.org/abs/2004.02653.

[174] Juan Zhao et al. “Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction.” In: *Scientific Reports* 9.1 (2019), pp. 1–10. ISSN: 20452322. DOI: 10.1038/s41598-018-36745-x. URL: http://dx.doi.org/10.1038/s41598-018-36745-x.

[175] Robert Chen et al. “Recurrent neural networks for early detection of heart failure from longitudinal electronic health record data: Implications for temporal modeling with respect to time before diagnosis, data density, data quantity and data type Access.” In: *Physiology \& behavior* 176.3 (2019), pp. 139–148. DOI: 10.1161/CIRCOUTCOMES.118.005114.Recurrent.

[176] Gregory Privitera. *Statistics for the Behavioral Sciences*. 1st. SAGE Publications, 2017, p. 816. ISBN: 1506386256.

[177] Bruce F Walker and Antony Ugoni. “THE t TEST: An Introduction.” In: *COMSIG Review* 4.2 (1995), pp. 37–40. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2050377/pdf/cr042-037b.pdf.

[178] Nan M Laird and James H Ware. “Random-Effects Models for Longitudinal Data Author (s): Nan M . Laird and James H . Ware Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2529876 REFERENCES Linked references are available on JSTOR for this article.” In: *Biometrics* 38.4 (1982), pp. 963–974.

[179] Edward Joseph Caruana et al. “Longitudinal studies.” In: *Journal of Thoracic Disease* 7.11 (2015), E537–E540. ISSN: 20776624. DOI: 10.3978/j.issn.2072-1439.2015.10.63.

[180] Kristin Gustavson et al. “Attrition and generalizability in longitudinal studies: Findings from a 15-year population-based study and a Monte Carlo simulation study.” In: *BMC Public Health* 12.1 (2012). ISSN: 14712458. DOI: 10.1186/1471-2458-12-918.