Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

# Adaptiv sequentielle Auswahl von Messzeitpunkten zur Vorhersage der mentalen Gesundheit

MASTER-ARBEIT

Juni 2025

Autor

Deseö, Florian

Begleitperson

Dr. Wyss, Patric

Praxispartner

Abteilung Klinische Psychologie und Psychotherapie, Universität Bern

Prof. Dr. Berger, Thomas

## Zusammenfassung

Die Zunahme psychischer Störungen führt zu Leiden der Betroffenen sowie zu diversen gesellschaftlichen Problemen. Wegen steigender Gesundheitskosten und den Vorteilen einer frühen Intervention gilt neben der präzisen Prognosefähigkeit auch eine möglichst frühe und kosteneffiziente Prognose zu erreichen. Dabei bieten sequentielle Entscheidungen die Möglichkeit, mit der adaptiven Auswahl von Messzeitpunkten dies zu berücksichtigen. In dieser Arbeit wurde ein explorativer, datengetriebener Ansatz zur Vorhersage zukünftiger depressiver Symptome verfolgt. Dabei wurde untersucht, inwiefern die schrittweise Auswahl von Messzeitpunkten eine präzise, kosteneffiziente und frühzeitige Prädiktion ermöglicht. Es wurden zwei bereits bestehende Datensätze untersucht mit Informationen zum Lebensstil, der Gesundheit, der Arbeit und biometrische Messungen. Durch zwei frühere Messzeitpunkte wurde vorhergesagt, ob an einem zukünftigen dritten Messzeitpunkt erhöhte depressive Symptomatik vorliegt (binäre Klassifikation). Dafür wurden Machine Learning Modelle mit automatisierter Variablenselektion und Diskriminanzanalysen genutzt. Es konnte gefunden werden, dass im Vergleich zu zwei starren Messzeitpunkten die adaptiv sequentielle Auswahl mit einer deutlichen Reduktion der Anzahl zweiter Messzeitpunkte eine gleich gute Genauigkeit (Balanced Accuracy) erreicht. Dabei können die durchschnittlichen Gesamtkosten (Summe aus Messkosten und Fehlklassifikationskosten) reduziert werden. Im Vergleich zu nur einem Messzeitpunkt konnte gefunden werden, dass bei niedrig definierten Messkosten die sequentielle Auswahl weniger durchschnittliche Gesamtkosten verursacht. Bei steigenden definierten Messkosten verursacht jedoch die sequentielle Auswahl teilweise leicht höhere durchschnittliche Gesamtkosten. Dies kann in der Praxis helfen Entscheidungsregeln zu definieren und Monitoring gezielt und kosteneffizient einzusetzen. Künftige Forschung sollte validierte psychologische Variablen verwenden und den Zeitabstand zwischen Messungen systematisch variieren.

*Schlagwörter:* adaptiv sequentielle Auswahl, depressive Symptomatik, Machine Learning

## Abstract

The increase in mental disorders leads to suffering for those affected and to challenges for society. Because of the importance of an early prognosis and growing costs in healthcare, the aim is to achieve a precise prediction in a short time with low costs with the use of sequential selection of measurement points. This paper analyzes two existing datasets using a data driven, exploratory approach. Variables on lifestyle, health, work and biometric measures were used. Two previous measurement points were used to predict whether a defined threshold of depressive symptoms would be exceeded at a future third measurement point (binary classification). Machine Learning Models with automated variable selection and discriminant analysis were used. Compared to two fixed measurement points for everyone, adaptive sequential selection could reach the same prediction quality, measured by balanced accuracy, with a reduction in measurements. The average total costs (sum of costs of misclassification and of measurement) were reduced as well. When compared to the fixed selection of one measurement point the mean total costs were lower as well when the measurement costs were defined as low but with increasing defined measurement costs, the adaptive sequential selection caused slightly higher mean total costs in some cases. These findings can help to define decision rules and can be used for the application of cost efficient targeted monitoring of depressive symptoms. Future research should focus on using psychologically validated measures in the adaptive sequential selection of measurement points and should vary the time intervals between measurement points.


*Keywords:* Adaptive sequential selection, depressive symptoms, Machine Learning

# Inhaltsverzeichnis

**1. Einleitung ..... 6**

**2. Theorie ..... 8**

2.1. Depressionen und Einflussfaktoren ..... 8

*2.1.1. Depressive Störungen. ..... 8*

*2.1.2. Selbstberichtete Fragen zum Lebensstil, Arbeitsleben, Freizeit, Gesundheit und Netzwerk. ..... 11*

*2.1.3. Biometrische Messungen. ..... 14*

2.2. Big Data und Machine Learning ..... 17

*2.2.1. Random Forest und Hyperparameter Tuning. ..... 20*

*2.2.2. Diskriminanzanalyse.. ..... 26*

*2.2.3. Sequentielle Selektion.. ..... 27*

2.3. Forschungsfragen ..... 29

**3. Methodik ..... 30**

3.1. Datensätze, Skalenbildung und Itemauswahl ..... 31

*3.1.1. Schweizer Haushalt-Panel. ..... 33*

*3.1.2. PSYCHE-D. ..... 39*

3.2. Statistische Modellierung ..... 43

*3.2.1. Übersicht des Vorgehens. ..... 43*

*3.2.2. Random Forests zur Implementierung der prädiktiven Modelle. ..... 45*

*3.2.3. Regressionsbasierte Diskriminanzmodelle. ..... 47*

*3.2.4. Sequentielle Auswahl von Messzeitpunkten. ..... 49*

**4. Ergebnisse ..... 52**

4.1. Prädiktionsmodelle ..... 52

*4.1.1. Schweizer Haushalt-Panel. ..... 52*

*4.1.2. PSYCHE-D. ..... 55*

4.2. Probabilistische Modelle nach Messzeitpunkt ..... 58

*4.2.1. Schweizer Haushalt-Panel. ..... 58*

*4.2.2. PSYCHE-D. ..... 60*

4.3. Sequentielle Selektion Prädiktionsleistung ..... 62

*4.3.1. Schweizer Haushalt-Panel. ..... 63*

*4.3.2. PSYCHE-D. ..... 65*

4.4. Sequentielle Selektion Kosten ..... 69

*4.4.1. Schweizer Haushalt-Panel. ..... 69*

*4.4.2. PSYCHE-D.* ..... *71*
**5. Diskussion** ..... **74**
5.1. Zusammenfassung ..... 75
*5.1.1. Prädiktiver Nutzen betrachteter Variablen gegenüber symptombasierter Vorhersage.* ..... *75*
*5.1.2. Beantwortung erste Forschungsfrage zur Prognosegüte und Frühzeitigkeit.* ..... *82*
*5.1.3. Beantwortung zweite Forschungsfrage zu Kosten.* ..... *84*
5.2. Limitationen ..... 88
5.3. Fazit und Ausblick ..... 95
Literaturverzeichnis ..... 97
Verzeichnis der verwendeten Hilfsmittel ..... 122
Tabellenverzeichnis ..... 123
Abbildungsverzeichnis ..... 124

# 1. Einleitung

Die zunehmende Belastung durch depressive Störungen führt zu Leiden der Betroffenen und zu diversen gesellschaftlichen Problemen (Bundesamt für Statistik, 2022; Lépine & Briley, 2011). So sind depressive Störungen unter anderem mit erhöhtem Suizidrisiko (20-fach) und kardiovaskulären Erkrankungen verbunden (Lépine & Briley, 2011). Als gesellschaftliche Probleme gelten unter anderem die steigenden Gesundheitskosten sowie die geringere Produktivität und Leistungsfähigkeit im Arbeitskontext (Bundesamt für Statistik, 2024; Lerner & Henke, 2008). Neben dem Ziel Gesundheitskosten gering zu halten, gilt auch eine möglichst frühzeitige Prädiktion von depressiven Symptomen als Ziel, da eine frühe Indikation und Behandlung für den Verlauf von Bedeutung sind (Kraus, Kadriu, Lanzenberger, Zarate Jr. & Kasper, 2019). Der Multi-Objective Ansatz versucht, diese verschiedenen Ziele auszubalancieren. Dabei ist das Ziel, möglichst früh, präzise und kosteneffizient die abhängige Variable vorherzusagen (Roijers, Vamplew, Whiteson & Dazeley, 2013).

Die sequentielle Auswahl bietet eine Möglichkeit, dies umzusetzen, wobei sich definierte Kosten für die Messung und die Fehlklassifikation als Einflussvariablen erfassen lassen und geprüft wird, ob mit einer geringeren Anzahl an sequentiell ausgewählten Messungen die Prognosegüte bewahrt und Kosten reduziert werden können (Denton, 2018; Zhang, Jeske, Li & Wong, 2016). Folglich finden sequentiell adaptive Entscheidungsregeln gerade im psychologischen und medizinischen Bereich bereits häufig Anwendung in der Forschung (Roddy, Pfammatter & Mayberry, 2023; Zhang & Saju, 2023). Bezüglich depressiven Störungen stand bislang jedoch vor allem im Fokus, inwiefern die individuell sequentiell adaptive Auswahl von unterschiedlichen Variablen die starre Auswahl dominieren kann (Devine et al., 2016; Loe, Stillwell & Gibbons, 2017; Tan, Cai, Li, Tan et al., 2018; Wong, Accortt, Choi & Bright, 2025). Dagegen ist das Ziel dieser Arbeit zu prüfen, inwiefern durch eine individuell zeitlich sequentielle Auswahl von Messzeitpunkten mit denselben

Variablen die Prognosegenauigkeit mit einer Verringerung der Anzahl Messungen und Kosten beibehalten werden kann bei der Vorhersage der künftigen depressiven Symptomatik. Dabei konnte die sequentielle Auswahl von Messzeitpunkten beispielsweise bei der Prognose einer Nierenfunktionsstörungen mit Blick auf Kosten und Prädiktionsgüte bereits erfolgreich angewendet werden (Zhang et al., 2016).

Um dies zu prüfen, wurden zwei bereits bestehende Datensätze verwendet (Makhmutova et al., 2021a; SHP Group, 2024). Der Fokus lag dabei einerseits auf Fragen zum Lebensstil, wie es das Diathese-Stress-Modell nahelegt (Zubin & Spring, 1977) und andererseits auf biometrischen Messungen wie Schritt- und Schlafdaten, die sich zur Prädiktion der mentalen Gesundheit eignen (Kim et al., 2019; Lovato & Gradisar, 2014; Mammen & Faulkner, 2013; Tsuno, Besset & Ritchie, 2005; Zhai, Zhang & Zhang, 2015). Ziel war es mittels Machine Learning Verfahren in einem datenbasierten, explorativen Vorgehen Muster in den Daten zu erkennen, die theoriebasiert nicht entdeckt worden wären (Adjerid & Kelley, 2018; Austin, McLane, Pieczkiewicz, Adam & Monsen, 2023; Elragal & Klischewski, 2017; Van Lissa, 2023).

In dieser Arbeit wird in einem ersten Schritt die relevante Theorie vorgestellt, wobei zuerst auf depressive Störungen eingegangen wird und dann erläutert wird, wieso die Prädiktorengruppen aus den vorhandenen Datensätzen ausgewählt wurden. Darauffolgend wird ein Überblick über die verwendeten Datensätze und das methodische sowie statistische Vorgehen in dieser Arbeit gegeben, bevor die Ergebnisse vorgestellt werden. Anschliessend folgt die Diskussion der Ergebnisse und des Vorgehens. Das Ziel dieser Arbeit ist es in einem explorativen Vorgehen zu prüfen, ob mit einer sequentiellen Auswahl von Messzeitpunkten weniger durchschnittliche Gesamtkosten und eine gleich gute Prognosegüte bei Reduktion der Anzahl Messungen erreicht werden kann im Vergleich zur starren Auswahl von einem oder zwei Messzeitpunkten.

# 2. Theorie

Im folgenden Teil der Arbeit werden grundlegende theoretische und klinische Konstrukte erläutert und definiert. Dabei wird in einem ersten Schritt die psychische Störung Depression vorgestellt und definiert, sowie empirisch und theoretisch belegte Einflussfaktoren des Auftretens dieser genannt. In einem weiteren Schritt werden die in dieser Arbeit verwendeten statistischen Verfahren theoretisch eingebettet sowie die Vor- und Nachteile dieser erläutert.

## 2.1. Depressionen und Einflussfaktoren

Nach der theoretischen Einbettung der depressiven Störung wird auf die Einflussfaktoren, die in der vorliegenden Arbeit zur Prognose verwendet wurden, genauer eingegangen. Beim Datensatz Schweizer Haushalt-Panel handelt es sich bei den Prädiktionsvariablen um selbstberichtete Fragen zu Lebensstil, Gesundheit (inklusive emotionalem Empfinden), Freizeit, Netzwerk und dem Arbeitsleben. Beim Datensatz PSYCHE-D standen hingegen neben einigen Fragen zum Lebensstil biometrische Messungen zur Schlafqualität und zu Schritten zur Verfügung.

**2.1.1. Depressive Störungen.** Die Depression ist eine Störung mit hoher gesellschaftlicher und wirtschaftlicher Bedeutung. So gab es in der Schweiz in den vergangenen Jahren einen starken Anstieg der Anzahl an Diagnosen (Bundesamt für Statistik, 2022). In der Schweiz sind gemäss aktuellsten Zahlen des Bundesamts für Statistik 7.9% der Männer und 11.7% der Frauen von depressiven Symptomen betroffen (Bundesamt für Statistik, 2022). Die World Health Organization (WHO) geht weltweit von einer Prävalenz von 4.4% aus, was zu über 300 Millionen betroffene Personen führt (WHO, 2017). Dabei wird jedoch auch darauf hingewiesen, dass die Prävalenz von depressiven Störungen stark abhängig ist vom Lebensstandard und soziodemographischen Aspekten wie Alter, Geschlecht und Nationalität (WHO, 2017). Die Depression ist somit eine häufige und ernstzunehmende

Störung, welche beeinflusst, wie sich eine Person fühlt, denkt und handelt (National Institute of Mental Health, 2024).

Gemäss der elften Auflage der International Classification of Diseases (ICD-11) der WHO sind depressive Störungen durch eine depressive Stimmung und Freudlosigkeit gekennzeichnet (WHO, 2024). Dabei treten weitere kognitive, verhaltensbezogene oder neurovegetative Symptome auf, welche die Funktionsfähigkeit einer Person erheblich beeinträchtigen (WHO, 2024). Die Depression ist somit eine komplexe Störung, welche diverse Ursachen, Risikofaktoren und Krankheitsbilder mit sich bringt (Strakowski, 2012). Besonders schwerwiegend ist dies, da das Suizidrisiko durch Depressionen deutlich erhöht wird (Angst, Angst & Stassen, 1999). Des Weiteren zeigt diese Definition auf, dass das emotionale Erleben einer Person ein wichtiger Prädiktor von depressiven Störungen ist (Morris, Bylsma & Rottenberg, 2009; WHO, 2024). Dabei ist die frühe Erkennung und Behandlung von depressiven Störungen ein wichtiges Anliegen, da somit die Symptome und beschriebenen Auswirkungen vermindert sowie der Behandlungserfolg gesteigert werden können (Kraus et al., 2019; Nazareth & Kendrick, 2014).

Wichtige wirtschaftliche Folgen der Zunahme an depressiven Störungen sind eine höhere Rate an Arbeitslosigkeit sowie Absentismus und verminderte Leistungsfähigkeit am Arbeitsplatz (Adler et al., 2006; Lerner & Henke, 2008). Des Weiteren spielen steigende Kosten im Gesundheitssystem zunehmend eine bedeutende gesellschaftliche Rolle. So stiegen die Gesundheitskosten in der Schweiz im Jahr 2022 auf 91.5 Milliarden Franken an, was einem Anstieg zum Vorjahr von 2.5% entspricht (Bundesamt für Statistik, 2024). Auch für Unternehmen spielen steigende Kosten durch die Zunahme an depressiven Störungen vermittelt durch mehr Absenzen und verminderte Leistung eine wichtige Rolle (Adler et al., 2006; Lerner & Henke, 2008). Dabei bieten adaptiv sequentielle Messungen auch im organisationalen Kontext eine interessante Option, kosteneffizient Risikofälle zu identifizieren. Dies ist besonders relevant, da gefunden wurde, dass Personen in der Schweiz

im Durchschnitt gleich gut mit steigenden Arbeitsanforderungen zurechtkommen, jedoch mehr Extremwerte bezüglich Erschöpfung und psychischer Störungen vorliegen (Gesundheitsförderung Schweiz, 2024). Dies verweist auf die Wichtigkeit von individuellen Massnahmen im organisationalen Kontext, was dadurch verstärkt wird, dass diese metaanalytisch betrachtet zu besseren Ergebnissen führen als generelle organisationale Massnahmen (Richardson & Rothstein, 2008). Die sequentiell adaptive Selektion von Messzeitpunkten würde dabei die Möglichkeit bieten, mit regelmässigem Monitoring und geringem Ressourcenaufwand (sprich reduzierter Anzahl Messzeitpunkte) frühzeitig und kosteneffizient die depressive Symptomatik vorherzusagen.

Dabei kann es relevant sein, statt eine zukünftige Diagnose vorherzusagen, zu betrachten, ob die zukünftige Symptomatik im kritischen Bereich liegt, da beispielsweise soziodemographische Aspekte auch einen Einfluss darauf haben, ob es zu einer Diagnose kommt (Williams, Chung & Muennig, 2017). Eine Möglichkeit zur Erfassung der Symptome bieten klinische Fragebogen wie der PHQ-9, welcher auf den DSM-IV-Kriterien für eine Diagnose basiert (Kroenke, Spitzer & Williams, 2001). Dabei ist jedoch zu beachten, dass der PHQ-9 Score die aktuellen depressiven Symptome misst und dies nicht unbedingt zur Prädiktion der künftigen depressiven Symptome verwendet werden kann, da diese variieren können und auch die Interpretation der Fragen einen Einfluss auf die Antworten hat (Kokanovic, Bendelow & Philip, 2013; Musliner, Munk-Olsen, Eaton & Zandi, 2016; Ong et al., 2022). Dies führt zur Empfehlung des multimodalen Prädiktionsansatzes, dass zur Prädiktion von Depressionen auch andere Variablen verwendet werden sollen (Kaplan, Roberts, Camacho & Coyne, 1987; Liu et al., 2019). Dazu gehören beispielsweise selbstberichtete Fragen zum Lebensstil, Arbeitsleben, Freizeit und Netzwerk sowie biometrische Messungen. Darauf wird im Folgenden genauer eingegangen. Dabei gilt zu beachten, dass bei der Vorhersage der depressiven Symptomatik gefunden werden konnte, dass neben der aktuellen depressiven Symptomatik oft nur wenig inkrementelle Varianz durch

zusätzliche Variablen erklärt werden kann, wenn ein geringer Zeitabstand zwischen den Messungen vorliegt (Andersson et al., 2023; Christodoulou et al., 2019; Keenan, Feng, Hipwell & Klostermann, 2009).

**2.1.2. Selbstberichtete Fragen zum Lebensstil, Arbeitsleben, Freizeit, Gesundheit und Netzwerk.** Selbstberichtete Fragen zum Lebensstil gehören zu wichtigen Prädiktoren der mentalen Gesundheit (Binnewies et al., 2021; Simjanoski et al., 2022; Tesarova, Pekacek & Porrovecchio, 2023). Dies kann unter anderem durch das Diathese-Stress-Modell erklärt werden (Zubin & Spring, 1977). Das Diathese-Stress-Modell ist ein integrativer Ansatz, welcher für das Entwickeln psychischer Störungen das Einwirken biologischer, psychologischer und sozialer Faktoren beschreibt (Hankin & Abela, 2005; Hoyer & Knappe, 2020). Dabei wird auf der einen Seite die Vulnerabilität einer Person genannt, die die Anfälligkeit beschreibt, welche eine Person hat, eine psychische Störung zu entwickeln. Diese besteht wiederum aus genetischen wie auch aus erlernten psychosozialen Komponenten (wie beispielsweise Persönlichkeit und Glaubenssätze wie Kontrollgefühl; Hakulinen et al., 2015; Hoyer & Knappe, 2020; Presson & Benassi, 1996). Auf der anderen Seite werden für das Entwickeln einer psychischen Störung neben der Vulnerabilität auch auslösende Ereignisse genannt, welche als Stressoren bezeichnet werden. Damit werden biologische, soziale und psychologische Anpassungsreaktionen gemeint, welche gezeigt werden müssen, um auf kritische Lebensereignisse oder Alltagssituationen reagieren zu können (Hankin & Abela, 2005; Hoyer & Knappe, 2020). Zuletzt werden noch Ressourcen beschrieben, die zur Resilienz beitragen und somit der Entwicklung einer depressiven Störung vorbeugen können (Höltge, Theron & Ungar, 2022). Ressourcen sind somit Protektivfaktoren, die eine Person vor einem negativen Verlauf schützen (Hankin & Abela, 2005; Hoyer & Knappe, 2020). So können beispielsweise Fragen zu erlebtem Stress am Arbeitsplatz und sozialer Unterstützung vermittelt über Stressoren und Ressourcen das zukünftige Vorhandensein einer Depression vorhersagen, wobei jedoch auch die Vulnerabilität von entscheidender Bedeutung ist

(beispielsweise Perfektionismus als Vulnerabilitätsfaktor; Colodro-Conde et al., 2018; Flett, Hewitt, Blankstein & Mosher, 1995).

Eine Weiterentwicklung des Diathese-Stress-Modells, welche spezifisch Stressoren und Ressourcen im Arbeitsleben betrachtet, ist das Job-Demands-Resources Modell, welches auf dem Job-Demand-Control Modell beruht (Demerouti, Bakker, Nachreiner & Schaufeli, 2001; Karasek, 1979). Das Job-Demand-Control Modell von Karasek (1979) brachte erstmals die Idee auf, dass stressreiche Situationen am Arbeitsplatz bei vorhandenen Ressourcen besser bewältigt werden können, als wenn keine Ressourcen vorhanden sind. Dabei wurde jedoch nur der Handlungsspielraum als Ressource betrachtet. Wenn vorhanden, führen gemäss dem Job-Demand-Control Modell Arbeitsstressoren zu aktiver, herausfordernder Arbeit. Wenn jedoch kein Handlungsspielraum vorhanden ist, führt das zu einer starken Stressreaktion bei hohen Arbeitsanforderungen (Karasek, 1979). In der Weiterentwicklung durch das Job-Demands-Resources Modell wurde festgehalten, dass jede Arbeitstätigkeit mit diversen Demands, also Stressoren, und Ressourcen einhergeht (Demerouti et al., 2001). Dabei gehen Stressoren mit Beanspruchung und Burnout einher, wohingegen Ressourcen mit Motivation, Wachstum und Engagement einhergehen (Demerouti et al., 2001). Demerouti et al. (2001) verweisen dabei auch auf Interaktionseffekte, wobei Ressourcen den Zusammenhang zwischen Demands und Beanspruchung verringern und Ressourcen vor allem bei hohen Anforderungen ihre positive Wirkung entfalten. Stand beim Job-Demands-Resources Modell ursprünglich das Syndrom Burnout als Folge von hohen Belastungen im Fokus, konnte das Modell auch die Entwicklung von depressiven Störungen vorhersagen (Hakanen, Schaufeli & Ahola, 2008; Santa Maria et al., 2018). Ein Kritikpunkt am Job-Demands-Resources Modell sind die fehlenden persönlichen Ressourcen für die Vorhersage der mentalen Gesundheit, weshalb in dieser Arbeit neben Variablen aus dem Berufsleben auch Variablen aus der Freizeit und dem Netzwerk betrachtet werden (Schaufeli & Taris, 2014). Eine Schnittstelle sind dabei Freizeit-Arbeits-Konflikte. Diese beschreiben, dass einer der Bereiche Freizeit oder Arbeit

den anderen zeitbasiert, verhaltensbasiert oder belastungsbasiert behindert, was unter anderem zu geringem psychologischem Wohlbefinden und geringer Arbeitszufriedenheit führt (Tsaur, Liang & Hsu, 2012; Tsaur & Yen, 2018).

Zu den genannten persönlichen Ressourcen neben dem Arbeitsleben zählen die Gesundheit und das soziale Netzwerk (Cacheda, Fernandez, Novoa & Carneiro, 2019; Cross-Denny & Robinson, 2017). Gesundheitliche Faktoren auf den Lebensstil bezogen (wie beispielsweise Übergewicht, übermässiger Alkoholkonsum und mangelnde Bewegung in der Freizeit), sowie auf das emotionale Erleben bezogen, gehören zu wichtigen Prädiktoren von depressiven Störungen (Cross-Denny & Robinson, 2017; Zhang, Woud, Becker & Margraf, 2018). Gemäss einer Metaanalyse von Holt-Lunstad, Smith, Baker, Harris und Stephenson (2015) gehen soziale Isolation und Einsamkeit jedoch mit noch höherem Sterberisiko einher als die genannten gesundheitlichen Faktoren Rauchen, Alkohol und Übergewicht. Dabei ist das soziale Eingebundensein und die Qualität der Beziehungen von entscheidender Bedeutung (Holt-Lunstad et al., 2015). Dies wird unter anderem durch das GENIAL (genomics-environment-vagus nerve-social interaction-allostatic regulation-longevity) Modell erklärt. Dieses beschreibt, dass soziale Beziehungen, durch den Vagusnerv auf physiologischer Ebene vermittelt, die Allostase positiv beeinflussen und somit zu Wohlbefinden und dem Ausbleiben von Krankheiten, wie beispielsweise Depressionen, führen (Kemp, Arias & Fisher, 2017). So konnte in einem systematischen Review belegt werden, dass wahrgenommene emotionale und instrumentelle Unterstützung sowie grosse, diverse Netzwerke der Entwicklung einer depressiven Störung entgegenwirken (Santini, Koyanagi, Tyrovolas, Mason & Haro, 2015).

Abschliessend lässt sich festhalten, dass es empirische und theoretische Belege gibt, dass Variablen aus dem Arbeitsleben, der Freizeit, dem sozialen Netzwerk, dem emotionalen Erleben, der Gesundheit und damit zusammenhängend dem Lebensstil die Entwicklung der mentalen Gesundheit und einer depressiven Störung vorhersagen können. Dazu kommen noch Lebensereignisse, wie beispielsweise der Tod einer nahestehenden Person, welche auch

wiederum mit Stressoren und Ressourcen im Zusammenhang mit dem Diathese-Stress-Modell einhergehen (Fried et al., 2015).

**2.1.3. Biometrische Messungen.** Wie bereits beschrieben zählen gesundheitliche Faktoren, auch auf den Lebensstil bezogen, zu wichtigen Prädiktoren der mentalen Gesundheit (Cross-Denny & Robinson, 2017; Zhang et al., 2018). Eine Möglichkeit biologische und gesundheitsbezogene Variablen zu erheben, ist mittels biometrischer Messungen. Dabei ist eine Möglichkeit, ohne grossen Aufwand Daten zu erheben, die der passiv gemessenen Daten (Onnela, 2021). Durch die digitale Phänotypisierung werden unter anderem biometrische Daten passiv erhoben, welche für die Diagnose und Behandlung psychischer Störungen verwendet werden können (Martinez-Martin, Greely & Cho, 2021). Dies wird jedoch im klinischen Rahmen der mentalen Gesundheit bislang nur wenig angewendet und die Vorteile der digitalen Phänotypisierung, darunter geringer Aufwand, längsschnittliche Daten, kontinuierliche Datenerhebung und Datenerhebung in realen Situationen, werden folglich erst selten genutzt (Liang, Zheng & Zeng, 2019; Onnela, 2021). So konnten beispielsweise De Angel et al. (2022) in einem Review zeigen, dass passiv durch Smartphone- und Wearable-Device-Sensoren gewonnene Daten das Auftreten einer Depression vorhersagen können (beispielsweise über Variablen Schlaf und Bewegung), diese Studien jedoch oft Probleme der Reproduzierbarkeit haben. Dies wird unter anderem durch kleine Stichprobengrössen, kurze Beobachtungszeiten und geringe Beschreibung der Stichprobenzusammensetzung erklärt (De Angel et al., 2022). Da beim untersuchten Datensatz PSYCHE-D neben Fragen zum Lebensstil nur die biometrischen Daten zu Schritten und Schlaf mit Wearables aufgezeichnet wurden, wird im Folgenden auf diese beiden Aspekte genauer eingegangen (Makhmutova et al., 2021b).

Das prädiktive Potenzial von Daten zu Schritten ist besonders auf die physische Aktivität zurückzuführen, welche einen starken Einfluss auf die Entwicklung depressiver Symptome hat (Kim et al., 2019; Mammen & Faulkner, 2013). Dabei gilt jedoch zu beachten, dass

metaanalytische Befunde darauf hindeuten, dass der Zusammenhang zwischen physischer Aktivität und psychischem Wohlbefinden kurvilinear ist (Pearce et al., 2022). Dabei gibt es mehrere Theorien, welche diesen Effekt erklären können. So wird beispielsweise aus biopsychologischer Forschung darauf hingewiesen, dass physische Aktivität Biomarker, welche mit depressiven Symptomen in Verbindung stehen, positiv beeinflussen kann (Archer, Josefsson & Lindwall, 2014). So reguliert gemäss Literaturreview von Archer et al. (2014) körperliche Aktivität die Hypothalamus-Hypophysen-Nebennierenrinde-Achse-Homöostase, steigert die Neuroplastizität durch erhöhte BDNF-Werte und beeinflusst den Monoamin-Stoffwechsel, indem Verfügbarkeit von Dopamin und Serotonin erhöht werden. Des Weiteren wird in der Endorphin-Hypothese postuliert, dass durch physische Aktivität Endorphine ausgeschüttet werden, welche das Wohlbefinden steigern (Hoffmann, 1997). Dabei gilt es jedoch zu beachten, dass auch zu viel Bewegung einen negativen Einfluss auf die Allostase haben kann und somit auf die Fähigkeit des Körpers auf Stress zu reagieren (Lehmann, Foster & Keul, 1998). Motivationspsychologische Theorien gehen hingegen stärker auf die Erfüllung von Bedürfnissen durch physische Aktivität ein. So kann Sport zur Erfüllung von den gemäss Selbstbestimmungstheorie von Deci und Ryan (1985) grundlegenden Bedürfnissen Autonomie, Kontrolle und sozialem Eingebundensein führen (Ryan, Williams, Patrick & Deci, 2009).

Auch biometrische Messungen zum Schlaf eignen sich zur Prognose von depressiven Störungen (Lovato & Gradisar, 2014; Tsuno et al., 2005; Zhai et al., 2015). Zum einen kann gestörter Schlaf zu Depressionen und weiteren psychischen Störungen führen (Becker, Langberg & Evans, 2015; Gregory et al., 2005; Whalen, Gilbert, Barch, Luby & Belden, 2017). Zum anderen gibt es neurobiologische Mechanismen, welche den Effekt zwischen Schlaf und Depression vermitteln. So konnte gefunden werden, dass das Melanin-konzentrierte Hormon an der Regulation von Schlaf und Stimmung beteiligt ist und Mikroinjektionen davon depressive Verhaltensweisen auslösen können (Torterolo et al.,

2015). Ein weiterer diskutierter Wirkmechanismus, der den Zusammenhang von Depressionen und Schlaf aufzeigt, ist wiederum die Hypothalamus-Hypophysen-Nebennierenrinde-Achse (HHNA), wobei hier kausale Belege noch fehlen (Asarnow, 2020; Rao, Hammen & Poland, 2009). Die Aktivität der HHNA wird wiederum häufig mit Stress in Verbindung gebracht (Gheshlagh et al., 2017; McEwen, 1998a; Pinel, Barnes & Pauli, 2018).

Folglich ist es relevant die Wirkung von Stressoren über die HHNA zu betrachten. Als Reaktion auf einen Stressor kommt es zu physiologischer Erregung und zur Ausschüttung vom adrenocorticotropen Hormon (ACTH) aus dem Hypophysenvorderlappen, was zur Freisetzung von Glucocorticoiden in der Nebennierenrinde führt. Dies verursacht wiederum diverse Stressfolgen wie die Freisetzung von Adrenalin und Noradrenalin im Nebennierenmark durch die sympathische Aktivierung, sowie dadurch die Aktivierung des Herz-Kreislauf-Systems, Unterdrückung der Sexualhormone und Beeinflussung des Immunsystems (Ehlert, 2014; Pinel et al., 2018). Dabei kommt es folgend zu einem Rückgang der physiologischen Stressreaktion, da die hohe Ausschüttung von Cortisol wiederum die ACTH-Sekretion senkt (Ramamoorthy & Cidlowski, 2016). Auf diesen Kreislauf geht das Allostase Modell ein. Dabei ist der Organismus darauf aus, sich an wechselnde Umweltbedingungen mit verschiedenen physiologischen Antworten anzupassen, was neben der adaptiven physiologischen Reaktion auch die Regeneration bedeutend macht (McEwen, 1998a). Wird der Zusatzaufwand durch nicht normale Stressreaktionen zu gross, kann es zu Erkrankungen, wie beispielsweise depressiven Störungen, kommen (Arnaldo, Corcoran, Friston & Ramstead, 2022). Dabei werden als nicht normale Stressreaktionen die fehlende Adaptation, die verlängerte Reaktion, die inadäquat tiefe Reaktion und wiederholte akute Stressreaktion von gewissen physiologischen Aspekten genannt (McEwen, 1998a). Dabei gilt zu beachten, dass hierzu häufig längere Zeitintervalle betrachtet werden (McEwen, 1998b). Dies verweist auf die Wichtigkeit der biometrischen Messungen für die Vorhersage der

mentalen Gesundheit und unterstreicht erneut die Wichtigkeit der Betrachtung von diversen Stressoren sowie Ressourcen.

Folglich lässt sich aus verschiedenen psychologischen Disziplinen normativ und evidenzbasiert ableiten, dass durch biometrische Messungen von Schritten und vom Schlaf depressive Symptomatik vorhergesagt werden kann. Dies ist besonders durch die berichteten Vorteile von biometrischen Messungen relevant. Dabei gilt jedoch zu beachten, dass Messungen mit Wearables unterschiedlich genau sein können und selten nach wissenschaftlichen Standards entwickelt und evaluiert werden (Peake, Kerr & Sullivan, 2018). Jedoch konnte gefunden werden, dass die Messungen bezüglich Schritt- und Schlafdaten mit Wearables eine Reliabilität von $r > .8$ erreichen können (Evenson, Goto & Furberg, 2015; Fuller et al., 2020).

## 2.2. Big Data und Machine Learning

Da der datengestützte Ansatz in der Forschung zunehmend Anwendung findet, wird bereits diskutiert, dass es sich um ein neues Zeitalter der Forschung handelt (Hey, Tansley & Tolle, 2009). Beim datengestützten Ansatz ist es wichtig, dass grosse Mengen an Daten zur Verfügung stehen, was als Big Data bezeichnet wird (Kitchin, 2014). Dabei ist gemäss Kitchin (2014) Big Data durch drei zentrale Aspekte definiert: Volume, Velocity und Variety. Volume beschreibt dabei die enorme Datenmenge, Velocity die Geschwindigkeit, mit der die Daten meist in Echtzeit erhoben werden und Variety die Vielfalt der Daten (strukturiert, semistrukturiert und unstrukturiert). Dabei sind Daten, welche bereits im Zahlenformat vorliegen, strukturiert und Daten, welche nicht im Zahlenformat vorliegen, wie beispielsweise Bilder oder Videos, unstrukturiert (Chen & Wojcik, 2016). Weitere Merkmale sind, dass Big Data oft möglichst umfassend ist statt auf Stichproben basierend, eine hohe Auflösung und eindeutige Identifizierbarkeit von Datenpunkten vorliegt und gemeinsame Felder enthalten

sind, die die Verknüpfung von Datensätzen/Datenpunkten ermöglichen und somit eine hohe Flexibilität gewährleisten (Kitchin, 2014).

So ist ein Vorteil des datengestützten Ansatzes und Big Data, dass durch Analysen von grossen Datenmengen Muster gefunden werden können, welche mit einem rein deduktiven Vorgehen nicht entdeckt worden wären (Adjerid & Kelley, 2018; Austin, et al., 2023; Elragal & Klischewski, 2017; Van Lissa, 2023). Ausserdem wird diskutiert, dass die interdisziplinäre Zusammenarbeit gefördert werden kann, da jeder mit statistischem Verständnis die Daten interpretieren kann, ohne die Notwendigkeit komplexe Theorien verstehen zu müssen (Kitchin, 2014). Weiter können die Daten für sich selber sprechen und werden weniger stark von einer erläuterten Theorie beeinflusst, wobei jedoch argumentiert wird, dass Daten immer unter einem gewissen Frame betrachtet werden (Kitchin, 2014). Weitere Vorteile sind, dass induktiv neue Theorien gebildet werden, viel Flexibilität bei der Auswertung vorhanden ist (im Gegensatz zum Hypothesentest) und mittels Machine Learning hohe prädiktive Validität erreicht werden kann (Van Lissa, 2023). So konnte auch beim Gesundheitsverhalten nachgewiesen werden, dass durch ein datenbasiertes Vorgehen Muster gefunden werden, die rein durch Theorie nicht gefunden worden wären (Austin et al., 2023). Um diese Vorteile erreichen zu können, ist es wichtig, dass zuerst definiert wird, welche Datensätze verwendet werden, die Inklusionskriterien definiert werden und bestimmt wird, wie mit fehlenden Werten umgegangen wird. Erst darauf folgend soll überlegt und begründet werden, wie ausgewertet wird (basierend auf Theorie der statistischen Verfahren) und abschliessend sollen die Ergebnisse interpretiert werden (Elragal & Klischewski, 2017).

Es werden einige Nachteile und Gefahren des datenbasierten Vorgehens mit Big Data diskutiert. So ist es wichtig, dass das Ziel der Datenanalyse bereits vor dem Suchen der Datensätze und folgender Auswertung klar und evidenzbasiert begründet ist (Elragal & Klischewski, 2017). Ausserdem sollen mögliche Probleme und Gefahren bekannt sein und beachtet werden. Darunter werden in der Literatur der Streetlight Effekt, Overfitting,

finanzielle Kosten der Daten, mangelnde Datenqualität und erschwerte Interpretierbarkeit diskutiert (Adjerid & Kelley, 2018; Dietterich, 1995; Elragal & Klischewski, 2017). Der Streetlight Effekt beschreibt die Gefahr, dass statt den wichtigsten Themen und Problemen nur Aspekte untersucht werden, bei denen eine grosse Menge an Daten vorliegt (Elragal & Klischewski, 2017). Bei der mangelnden Datenqualität gilt besonders zu beachten, dass Big Data oft nicht nach wissenschaftlichen Standards gewonnen wird und keine Gütekriterien, wie Reliabilität, Validität und Objektivität, angegeben werden (Elragal & Klischewski, 2017). Ein Problem, das besonders bei der klinisch psychologischen Datenerhebung auftritt, ist, dass es bei psychischen Störungen oft zu fehlenden Werten kommt (de Winter et al., 2005). Dies ist besonders problematisch, da es sich dabei häufig um einen systematischen Stichprobenausfall handelt (Chang, Yang, Tang & Ganguli, 2009). Eine mögliche Lösung ist dabei die Datenimputation (Andridge & Little, 2010). Als weitere Gefahren werden die ethischen Fragen bezüglich Datenschutz genannt und die Komplexität der statistischen Auswertungen, für welche oft Machine Learning Modelle verwendet werden (Adjerid & Kelley, 2018).

Machine Learning Modelle können automatisiert Wissen generieren, Algorithmen trainieren sowie Muster identifizieren und dabei auch schwache Zusammenhänge erkennen und prognostisch nutzen (Schmidt, 2022). Dabei lässt sich Supervised und Unsupervised Machine Learning unterscheiden. Währenddem beim Supervised Machine Learning eine klare abhängige Variable definiert wird, wird beim Unsupervised Machine Learning keine abhängige Variable definiert (Berry, Mohamed & Yap, 2019). Eine grosse Gefahr des Machine Learnings ist das Overfitting, welches beschreibt, dass Machine Learning Modelle die Muster in den Daten so genau replizieren, dass die Modelle andere Daten, welche nicht für die Erstellung des Modells verwendet wurden, ungenau vorhersagen (Dietterich, 1995). Um durch Overfitting zu optimistische Modellbeurteilungen zu vermeiden, können die Daten entweder in Test- und Trainingsdaten unterteilt oder k-Fold Cross Validation verwendet werden, wobei letzteres zu einem geringeren Datenverlust bei der Modellschätzung und

Beurteilung führt (Dietterich, 1995). Bei der k-Fold Cross Validation werden die Daten in *k* Folds unterteilt und das Modell wird jeweils anhand von $k - 1$ Folds erstellt und am verbleibenden Fold evaluiert, was *k*-Mal wiederholt wird (James, Witten, Hastie & Tibshirani, 2017).

Der datengestützte Ansatz konnte sich in der (induktiven) Forschung bereits etablieren und auch schon zur Prognose von depressiven Störungen verwendet werden. So konnte gezeigt werden, dass mittels Machine Learning Modellen eine präzise Prognose von depressiven Störungen erzielt werden kann (Cacheda et al., 2019). Jedoch werden Modelle aus biometrischen Messungen und Machine Learning in der Praxis bislang erst wenig angewendet (Eder et al., 2023). Diese Arbeit kommt somit einem Call for Action nach, entsprechende Modelle zu entwickeln, die in Zukunft zu einer praktischen Anwendung führen können (Eder et al., 2023; Liang et al., 2019).

**2.2.1. Random Forest und Hyperparameter Tuning.** Random Forest ist ein nichtparametrisches Ensemble-Lernverfahren, das aus einer Vielzahl an Entscheidungsbäumen besteht. Im Gegensatz zu einzelnen Entscheidungsbäumen werden bei jedem Split nicht alle verfügbaren Variablen betrachtet, sondern eine zufällige Teilmenge davon (Breiman, 2001). Anschliessend wird diejenige Variable ausgewählt, die die beste Trennung ermöglicht. Diese Methode reduziert Overfitting und verbessert die Modellleistung (Ali, Khan, Ahmad & Maqsood, 2012; Breiman, 2001). Die fallbezogene Zuteilung zu einem Prognosewert erfolgt über den Durchschnitt von der Zuteilung aller erstellen Entscheidungsbäume (Breiman, 2001). Da bei Random Forests eine Vielzahl an Entscheidungsbäumen erstellt werden und die genaue Entscheidungsbasis nur schwer erkennbar ist, handelt es sich im Gegensatz zu einfachen Entscheidungsbäumen um ein Black Box Verfahren (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo & Rigol-Sanchez, 2012). Verschiedene Vergleichsstudien konnten zeigen, dass sich Random Forests im Vergleich zu anderen Machine Learning Verfahren besser zur Prädiktion von Depressionen und anderen

psychischen Störungen eignen (Cho, Yim, Choi, Ko & Lee, 2019; Chung & Teo, 2022; Sekhar, Minal & Madhu, 2016; Shin, Lee, Adeluwa & Hur, 2020; Wanderley Espinola, Gomes, Mônica Silva Pereira & dos Santos, 2022). Dabei wurde im Review von Cho et al. (2019) kritisiert, dass nur selten in wissenschaftlichen Arbeiten erläutert wird, wieso ein Machine Learning Verfahren ausgewählt wurde und was die Vor- und Nachteile dieser Wahl sind. Im Folgenden wird daher die Wahl des Vorgehens in dieser Arbeit theoriegeleitet und evidenzbasiert hergeleitet.

Random Forest Verfahren eignen sich, wenn eine hohe Genauigkeit wichtig ist, komplexe Datensätze verwendet werden, eine Klassifizierung gemacht wird, nichtlineare Zusammenhänge vorliegen, die Daten Ausreisser enthalten könnten und die Zusammenhänge der Prädiktionsvariablen komplex sind (Biau & Scornet, 2016; James et al., 2017). Die flexible Modellierung der Zusammenhänge ist besonders wichtig, da im Kapitel 2.1. darauf verwiesen wurde, dass viele Effekte bezüglich depressiver Symptomatik voneinander abhängig und teilweise kurvilinear sind. Hingegen eignen sich Random Forests weniger für Untersuchungen, wenn eine hohe Interpretierbarkeit angestrebt wird und nur wenig Zeit für die Berechnung der Modelle zur Verfügung steht (Biau & Scornet, 2016). Dabei kann die Wichtigkeit der Variablen für die Prädiktion über die Variablenwichtigkeit veranschaulicht werden (Biau & Scornet, 2016).

Ausserdem wird als weiterer Vorteil diskutiert, dass im Vergleich zu anderen Machine Learning Modellen nur wenige Hyperparameter existieren, die bei Random Forests einen Einfluss auf die Ergebnisse haben können (Biau & Scornet, 2016). Hyperparameter sind Einstellungen oder Konfigurationen der Modelle, die nicht von den Daten geschätzt werden und vor der Erstellung der Modelle gesetzt werden müssen (Bartz, 2023). Diese sind innerhalb eines gewählten Bereichs frei wählbar und beeinflussen die Erstellung des Modells und somit auch die Modelleistung (Bartz, 2023). Die wichtigsten Hyperparameter bei Random Forests sind die Anzahl der Bäume, die Anzahl der Variablen, die bei jedem Split

zufällig ausgewählt werden, die Regel für die Entscheidung der Splits und die mindeste Anzahl Fälle pro Knoten (wobei beachtet werden muss, dass je nach Implementierung der Random Forests die Bezeichnung und Anwendung der Hyperparameter variieren kann; Bartz, 2023). Da es wenig theoretische Basis für die Default-Werte gibt, lohnt es sich, bei den Hyperparametern verschiedene systematisch auszuprobieren, da sich die Default-Werte oft nicht eignen für das Ziel der Untersuchung (Rimal, Sharma & Alsadoon, 2024; Scornet, 2017; Valavi, Elith, Lahoz-Monfort & Guillera-Arroita, 2021).

Bei der Anzahl an Entscheidungsbäumen wird empfohlen, 500 zu wählen, jedoch sollte dies mit der Out-of-Bag Error (OOB) Kurve für die vorliegenden Daten verifiziert werden (Bartz, 2023; Probst, Wright & Boulesteix, 2019; Scornet, 2017). Generell gilt, dass mehr Entscheidungsbäume die Stabilität erhöhen und den Generalisierungsfehler minimieren, jedoch sollte auch die Rechenzeit beachtet werden, welche linear mit der Anzahl der Bäume ansteigt (Breiman, 2001; Scornet, 2017).

Bei der Anzahl der Variablen, die bei jedem Split zufällig ausgewählt werden, gilt, dass kleine Werte oft präzisere Ergebnisse liefern, grosse jedoch stabiler sind (Probst et al., 2019). Auch hier gibt es keinen theoretischen Wert, der verwendet werden sollte und es empfiehlt sich, mehrere Werte auszuprobieren, da der beste Wert beispielsweise auch von der Anzahl relevanter Variablen abhängig ist (Breiman, 2001; Probst et al., 2019; Scornet, 2017). Als Empfehlung wird oft der Wert $\sqrt{P}$ genannt, jedoch basiert dieser auf Simulationen und nicht auf theoretischen Annahmen (Scornet, 2017).

Bei der Regel für die Entscheidung der Splits konnte noch keine gefunden werden, die konstant besser abschneidet als andere, jedoch konnte gezeigt werden, dass die Wahl der Entscheidungsregel auch einen Einfluss auf die Modelleistung haben kann (Probst et al., 2019). Der Default-Wert Gini präferiert unabhängige Variablen, die metrisch sind und viele Kategorien haben und hat Probleme mit unausgewogenen Daten, was generell eine Herausforderung von Random Forests und Machine Learning darstellt (Provost, 2000; Valavi

et al., 2021). Dabei wird bei der Entscheidungsregel Gini jeweils betrachtet, wie die Gini-Impurity, sprich die Unreinheit der Knoten bei Klassifizierungen, am stärksten vermindert werden kann (Breiman, 2001; Probst et al., 2019). Die Entscheidungsregel Extratrees ermöglicht durch zufällige Wahl der Splits eine robustere Leistung und schnellere Rechenzeit, was vor allem bei grossen und komplexen Datensätzen wichtig ist (Geurts, Ernst & Wehenkel, 2006). Dabei wird bei jeder Variable, die beim entsprechenden Split zufällig ausgewählt wurde, ein zufälliger Splitpunkt gewählt und folgend die Variable verwendet, die durch die zufällige Auswahl die Gini-Impurity am stärksten reduzieren kann (Geurts et al., 2006). Die Entscheidungsregel Hellinger wird vor allem bei ungleich verteilten Klassen empfohlen, da sie unempfindlich gegenüber Klassenungleichgewichten ist und eine gleichmässige Berücksichtigung der Klassen ermöglicht (Cieslak & Chawla, 2008; Valavi et al., 2021). Dabei wird jeweils der Split ausgewählt, der zur grössten Hellinger-Distanz führt, welche ein Mass der Trennung zwischen zwei Wahrscheinlichkeitsverteilungen ist (Aler, Valls & Boström, 2020).

Der Hyperparameter mindeste Anzahl Fälle pro Knoten ist im Default eins, was zu tiefen Bäumen und somit Overfitting führen kann und besonders bei grossen Datensätzen zu klein ist (Probst et al., 2019). Es wird empfohlen, den Wert grösser als eins zu wählen (Barreñada, Dhiman, Timmerman, Boulesteix & Van Calster, 2024). Auch hier gibt es keine theoretische Norm und es empfiehlt sich gerade bei ungleich verteilten Daten eine grössere Spannbreite beim Hyperparametertuning zu verwenden (Probst et al., 2019; Scornet, 2017; Valavi et al., 2021). Ausserdem kann Overfitting bei Random Forests entgegengewirkt werden, indem stark korrelierte Variablen ausgeschlossen werden und die Anzahl der Entscheidungsbäume reduziert wird (Barreñada et al., 2024). Dabei argumentieren Barreñada et al. (2024), dass Overfitting an die Trainingsdaten bei Random Forests weniger problematisch ist, solange Testdaten, welche nicht zur Erstellung des Modells verwendet wurden, präzise vorhergesagt werden können.

Da die Hyperparameter auch voneinander abhängig sind, wird empfohlen, dass die Hyperparameter in Form eines Tuning Rasters in allen möglichen Kombinationen evaluiert werden (Geurts et al., 2006). Für die Auswahl der besten Hyperparameterkombination sollte bei unausgewogenen Daten ein Mass verwendet werden, welches diese ungleiche Verteilung berücksichtigt. Geeignet sind beispielsweise Area Under the Precision-Recall Curve (PRAUC) oder Balanced Accuracy (BA; Hancock, Khoshgoftaar & Johnson, 2023; Liu et al., 2023). Der PRAUC gibt die Fläche unter der Precision-Recall Kurve an, welche den Zusammenhang zwischen Präzision und Sensitivität modelliert (Davis & Goadrich., 2006). Die BA ergibt sich aus der Summe von Sensitivität und Spezifität geteilt durch zwei und entspricht somit dem ungewichteten Mittelwert dieser Werte, wobei auch eine Gewichtung vorgenommen werden kann, wenn eine der Klassen von grösserer Bedeutung ist (Brodersen, Ong, Stephan & Buhmann, 2010). Um dabei Overfitting zu vermeiden, empfiehlt sich nested Cross Validation (Wainer & Cawley, 2021). Dabei wird der Datensatz in $k_{\text{innere}}$ und $k_{\text{äussere}}$ Folds ($k$ muss nicht identisch sein) unterteilt, wobei für $k$ ein Wert zwischen fünf und zehn empfohlen wird (Hastie, Tibshirani, Friedman & Franklin, 2017). Dies erlaubt eine Anwendung auf Daten, die weder bei der Auswahl der Hyperparameter noch der Erstellung des Modells beteiligt waren und folgend zur abschliessenden Beurteilung der Prognosegüte verwendet werden (Schwyck, Lieberman & Parkinson, 2020). Im Gegensatz steht die flat Cross Validation, welche den Datensatz in nur innere Folds unterteilt und die Prognosegüte sowie die Hyperparameter an denselben Folds schätzt (Wainer & Cawley, 2021). Dabei gilt, dass die nested Cross Validation zwar rechenaufwendiger ist, dafür die Prognosegenauigkeit realistischer und unverzerrt geschätzt werden kann (Berrar, 2018; Wainer & Cawley, 2021). Dabei wird jedoch aus praktischer Sicht darauf hingewiesen, dass die weniger rechenintensive flat Cross Validation in vielen Fällen zu einer vergleichbaren Modellwahl und Prädiktionsgüte führt (Wainer & Cawley, 2021).

Ein weiteres Problem von Random Forests ist, dass diese Schwierigkeiten bei der Prädiktionsleistung aufweisen, wenn eine grosse Anzahl an Variablen übergeben wird von denen sich viele nicht zur Prädiktion eignen (Hastie et al., 2017). Aus diesem Grund empfiehlt sich eine automatisierte Variablenselektion. Dies ist ausserdem wichtig, da diskutiert wird, dass für jede verwendete Variable bei Random Forests zwischen 20 und 50 Beobachtungen vorliegen sollten, wobei es auch bei mehr als 200 Beobachtungen pro Variable zu instabilen Schätzungen kommen kann (van der Ploeg, Austin & Steyerberg, 2014). Zwar sind Random Forests vergleichsweise robust gegenüber hoher Dimensionalität, problematisch wird es jedoch, wenn viele Prädiktoren und eine geringe Stichprobengrösse vorhanden sind (Biau & Scornet, 2016). Zwar besteht die Möglichkeit, die Wichtigkeit einer Variable über die Variablenwichtigkeit zu bestimmen, jedoch sollte bei der Variablenselektion nicht nur diese zur Entscheidung verwendet werden (Strobl, Boulesteix, Zeileis & Hothorn, 2007). Ausserdem konnte gezeigt werden, dass durch eine automatisierte Variablenselektion bessere Modelleistungen bei der Vorhersage von Depressionen mit weniger Variablen erzielt werden als bei Variablenauswahl durch Expert*innen (Zhang, Liu, Silenzio, Qiu & Gong, 2020). Bei einer binären, unausgewogenen abhängigen Variable schneidet gemäss verschiedenen Studien Boruta zur Variablenselektion am besten ab und es sollten keine Backwards Selektion Verfahren verwendet werden (Degenhardt, Seifert & Szymczak, 2019; Fox et al., 2017; Speiser, Miller, Tooze & Ip, 2019).

Bei der Variablenselektion mit Boruta werden zufällige Schattenvariablen erzeugt, indem die ursprünglichen Variablen permutiert werden. In wiederholten Random Forest Läufen (auf 100 festgelegt) wird die prognostische Wichtigkeit (Variablenwichtigkeit) jeder echten Variable mit dem Maximum der Schattenvariable verglichen (Degenhardt et al., 2019; Kursa & Rudnicki, 2010). Folgend werden unsichere Zuteilungen zugewiesen, indem der Median der prognostischen Wichtigkeit der Schattenvariable mit dem der entsprechenden Prädiktiobsvariable verglichen wird (Kursa & Rudnicki, 2010). Die Variablenauswahl ist zwar

auch von den Hyperparametern der Random Forests abhängig, Boruta ist in der Hinsicht jedoch robust, was darauf hindeutet, dass die Variablenauswahl vor dem Hyperparametertuning erfolgen kann (Kursa & Rudnicki, 2010).

**2.2.2. Diskriminanzanalyse.** Die Diskriminanzanalyse ist ein wichtiges parametrisches Verfahren zur Klassifikation und wird häufig in der Medizin und Psychologie verwendet, um Vorhersagen zu Krankheitsverläufen zu machen (Huberty, 1975; Johnston, 2023). Es kann zwischen linearer und quadratischer Diskriminanzanalyse unterschieden werden (Ghojogh & Crowley, 2019; James et al., 2017). Beide setzen normalverteilte Werte voraus, jedoch geht die lineare Diskriminanzanalyse von der gleichen Kovarianzmatrix in beiden Gruppen aus (homogene Streuung), währenddem die quadratische Diskriminanzanalyse zulässt, dass sich die Kovarianzmatrixen zwischen den Klassen unterscheiden. Dadurch entstehen bei der linearen Diskriminanzanalyse lineare Entscheidungsgrenzen und bei der quadratischen Diskriminanzanalyse Quadratische, welche flexibler, jedoch auch komplexer sind (Ghojogh & Crowley, 2019). Beide Verfahren basieren auf der Wahrscheinlichkeitsverteilung der Daten und versuchen die Vorhersagepräzision zu maximieren (Ghojogh & Crowley, 2019). Dabei kann im Satz von Bayes zwischen a-priori und a-posteriori Wahrscheinlichkeit unterschieden werden (Hastie et al., 2017). Die a-priori Wahrscheinlichkeit gibt an, wie wahrscheinlich eine Person zu einer Gruppe gehört, bevor die prädiktiven Daten zur Prognose dazu genommen werden (bei einer Prävalenz von 10% entspricht die a-priori Wahrscheinlichkeit ohne weitere Informationen beispielsweise 10%; Ramayah, Ahmad, Halim, Zainal, Lo, 2010). Diese kann auch auf Basis eines Prävalenzmodells beispielsweise mit soziodemografischen Informationen individuell verfeinert werden, um die unterschiedlichen Prävalenzen, beispielsweise nach Geschlecht, zu berücksichtigen (WHO, 2017). Die a-posteriori Wahrscheinlichkeit entspricht der Wahrscheinlichkeit einer Klassenzuordnung, die sich aus der a-priori Wahrscheinlichkeit und den erhobenen prädiktiven Variablen ergibt (Ramayah et al., 2010).

Eine mögliche Gefahr, die sich daraus ergibt, ist, dass die Daten nicht normalverteilt sind und somit die Annahme nicht erfüllen, was zu einer deutlich geringeren Klassifikationsleistung führt (Ghojogh & Crowley, 2019). Sind die Daten nicht normalverteilt, kann stattdessen die logistische Regression verwendet werden, wobei die Diskriminanzanalyse als mächtiger gilt, da sie die Wahrscheinlichkeitsverteilung der Gruppe berücksichtigt (Ramayah et al., 2010). Dabei kann zwischen diskriminanter und generativer Klassifikation unterschieden werden. Während bei der diskriminanten Klassifizierung (wie der logistischen Regression) die abhängige Variable gegeben der Prädiktoren modelliert wird, wird bei generativen Klassifizierungen zusätzlich die Verteilung der Prädiktoren in beiden Klassen für die Zuteilung betrachtet (Bernardo et al., 2007; James et al., 2023).

Des Weiteren kann es bei der Diskriminanzanalyse zu Problemen kommen, wenn die Prädiktoren hoch korreliert oder hochdimensional sind, weswegen es sich empfiehlt, die Diskriminanzanalyse bei einer geringen Anzahl an Prädiktoren anzuwenden (Ghojogh & Crowley, 2019). Ausserdem besteht bei der quadratischen Diskriminanzanalyse bei kleinen Datensätzen die Gefahr von Overfitting (Ramayah et al., 2010).

**2.2.3. Sequentielle Selektion.** Die sequentielle Selektion ist darauf ausgelegt, zeitlich oder inhaltlich unterschiedliche Messungen adaptiv auszuwählen. So wird beispielsweise mit dem Wissen aus der ersten Messung entschieden, ob eine weitere Messung gemacht wird, in Abhängigkeit davon, wie sicher die Klassifizierung bereits durch die erste Messung ist (Denton, 2018; Zhang et al., 2016). Im Gegensatz zu klassischen Machine Learning Modellen, wie beispielsweise dem Random Forest, die alle verfügbaren Daten gleichzeitig zur Klassifikation verwenden, werden bei der sequentiellen Selektion adaptive Entscheidungen getroffen. Dabei wird zunächst nur eine Teilmenge der Informationen genutzt und schrittweise entschieden, ob weitere Informationen benötigt werden. Ein Merkmal der in dieser Arbeit verwendeten sequentiellen Selektion mit Prospective Sequential Diagnosis with Neutral zones (POSEIDON) ist die Einführung einer neutralen Zone. Falls die Unsicherheit zu gross ist,

wird eine Entscheidung aufgeschoben und weitere Daten werden gesammelt, um die prognostische Sicherheit zu erhöhen (Wyss et al., 2023). Dabei haben auch die definierten Kosten der Messung und die definierten Kosten einer Fehlklassifikation (falsch positiv oder falsch negativ) einen Einfluss auf die zusätzliche Auswahl einer Messung und somit auf die Ergebnisse (Denton, 2018). Sind beispielsweise die Kosten einer Messung tief und die Kosten der Fehlklassifikation hoch, wird eher eine zweite Messung durchgeführt, als wenn die Kosten für die Messung hoch und die der Fehlklassifikation tief sind. Somit kann mit der Wahl der Kosten Einfluss auf die Auswahl der weiteren Messungen genommen werden, was eine Anpassung an den Kontext in der Praxis ermöglicht.

Das Ziel der sequentiellen Selektion ist es im Rahmen des Multi-Objective Ansatzes möglichst früh, präzise und kosteneffizient die abhängige Variable vorherzusagen (Roijers et al., 2013). So kommen sequentielle Entscheidungsstrategien und Studien vermehrt zum Einsatz im Bereich medizinischer Diagnosen (Poole et al., 2020; Trapeznikov & Saligrama, 2013; Wyss et al., 2023). Es konnten auch schon bezüglich der Prädiktion von depressiven Störungen die Vorteile der sequentiellen Auswahl belegt werden. In der Arbeit von Poole et al. (2020) wurde beispielsweise ein zweistufiges Screening Verfahren mit dem PHQ-2 und dem PHQ-8 validiert. Es konnte spezifisch für Flüchtlinge gezeigt werden, dass mit einer 25-prozentigen Reduktion der Personen, die den PHQ-8 ausfüllten, die Genauigkeit der Klassifikation nahezu gleich blieb (Poole et al., 2020). Diese Ergebnisse konnten in weiteren Studien bestätigt werden, in denen die sequentiell adaptive Auswahl von Variablen untersucht wurde (Devine et al., 2016; Loe et al., 2017; Tan et al., 2018; Wong et al., 2025). Dem Autor ist jedoch keine Studie bekannt, welche die zeitlich sequentielle Auswahl von Messzeitpunkten zur Vorhersage der depressiven Symptomatik mit fest ausgewählten biometrischen Variablen und Variablen zum Lebensstil untersucht.

Neben den aufgezeigten Chancen gibt es jedoch auch bei sequentiellen Entscheidungsregeln einige Nachteile und Risiken. So argumentiert Denton (2018), dass

benötigte longitudinale Messungen oft unvollständige Messungen enthalten, was zu geringerer Robustheit führen kann, es zu systematischen Fehlern kommen kann, wenn bei den Trainingsdaten gewisse Gruppen über- oder unterrepräsentiert sind und die Modelle auf validierten Fällen trainiert werden, währendem nicht ausgewertete Fälle nicht oder fehlerhaft validierte Fälle fälschlicherweise berücksichtigt werden. Ausserdem wird diskutiert, dass gewisse Aspekte, wie die definierten Kosten, teilweise unterschiedlich gesetzt werden können, was besonders die praktische Anwendung erschwert (Denton, 2018; Wyss et al., 2023). Zuletzt sind Beispiele der Übertragung in die Praxis eher selten, was auch teilweise auf das fehlende Vertrauen in der Praxis in die Modelle zurückzuführen sein kann (Denton, 2018).

## 2.3. Forschungsfragen

Da das Ziel dieser Arbeit nicht darin besteht, Theorien und damit in Verbindung stehende Hypothesen zu prüfen und zu testen, wird keine konfirmatorische, sondern eine datengetriebene, explorative Perspektive verfolgt. Für Vorhersagen werden Variablen und deren zeitlich mehrmaligen Messungen verwendet, welche sich als prädiktiv für zukünftige depressive Symptome erwiesen haben. Für ein prognostisches Modell wurden Vorhersagen mit unterschiedlichen sequentiell gewählten oder fixen Messsequenzen vorgenommen und evaluiert. Damit soll ein Proof of Concept dafür geliefert werden, dass eine sequentielle Auswahl effizienter ist als eine starre Auswahl von Messzeitpunkten, da die Ressourcen auf die schwierigen Fälle konzentriert werden. Entsprechend sollen die folgenden Fragestellungen untersucht werden:

1. Kann durch eine adaptive, sequentielle Auswahl von longitudinalen Messsequenzen (online messbare Variablen) die mentale Gesundheit präzise und frühzeitig vorhergesagt werden?

2. Kann mit einer sequentiellen Auswahl von Messungen die mentale Gesundheit kosteneffizienter vorhergesagt werden als mit einer fixen Auswahl von Messungen, die gleich ist für alle Personen?

Die Kosten ergeben sich dabei durch die Kosten der Fehlklassifikation und der Messungen. Als Mass für die Prädiktionsgüte wird auf Grund der ungleich verteilten Klassen die BA verwendet.

# 3. Methodik

In diesem Teil der Arbeit werden in einem ersten Schritt die beiden verwendeten Datensätze vorgestellt, wobei auch auf die Operationalisierung der abhängigen Variable (binäre Klassifikation) eingegangen wird. Dabei ist zu beachten, dass anders als ursprünglich geplant und im Exposé beschrieben nicht untersucht wurde, inwiefern Personen identifiziert werden können, die zu den Prognosezeitpunkten unter dem gewählten Threshold waren und dies auch stabil geblieben sind (Stables) oder schlussendlich darüber lagen (Converters). Stattdessen wurde untersucht, inwiefern vorhergesagt werden kann, ob die Personen in Zukunft über dem Threshold liegen. Es wurde also nicht untersucht, ob die Entwicklung erhöhter depressiver Symptome vorhergesagt werden kann, sondern, ob das zukünftige Vorhandensein dieser vorhergesagt werden kann. Um dabei für den aktuellen Wert der depressiven Symptomatik zum Prognosezeitpunkt zu kontrollieren, wurden zum Vergleich auch Modelle erstellt, welche nur die vorherigen Messungen der Variable, welche zur Operationalisierung der abhängigen Variable verwendet wurde, zur Prognose verwendet haben. Diese Abweichung im Vergleich zum Exposé wurde vorgenommen, da die Unterteilung in Stables und Converters zu einer geringen Anzahl Converters geführt hat, was zu instabilen Ergebnissen führt. Die Ergebnisse der Modelle mit Stables und Converters sind der Vollständigkeitshalber im Anhang A und B zu finden.

Bei der Datensatzerstellung und der Operationalisierung der abhängigen Variablen gibt es Unterschiede zwischen den Datensätzen PSYCHE-D und Schweizer Haushalt-Panel.

Darauffolgend ist das weitere Vorgehen für beide Datensätze dasselbe. Zuerst wird auf die Datensatzerstellung und Operationalisierung der abhängigen Variable eingegangen. Danach wird ein kurzer Überblick des weiteren Vorgehens gegeben bevor genauer auf die Erstellung der prädiktiven Modelle mit Random Forests eingegangen wird und anschliessend auf die regressionsbasierten Diskriminanzmodelle für eine sequentielle Auswahl von Messzeitpunkten.

## 3.1. Datensätze, Skalenbildung und Itemauswahl

Bei der vorliegenden Arbeit wurde der Fokus auf Datensätze gelegt, welche längsschnittliche Daten erfassen sowie Variablen zur depressiven Symptomatik und Variablen, welche sich theoretisch und empirisch hergeleitet zur Vorhersage der mentalen Gesundheit eignen. Aus Ressourcengründen wurde keine Datenimputation durchgeführt. In Tabelle 1 ist als Überblick zu erkennen, wie die abhängige Variable operationalisiert wurde. Es handelt sich bei der abhängigen Variable um eine binäre Klassifikation, ob eine erhöhte depressive Symptomatik in Zukunft vorhanden ist.

Tabelle 1

*Operationalisierung der abhängigen Variablen für zukünftige erhöhte depressive Symptomatik*

| Datensatz | Verwendete Variablen | Bedeutung (Range) | Entscheidung/Gewichtung | Anteil später erhöhte Symptomatik | Sensitivität | Spezifität |
|---|---|---|---|---|---|---|
| SHP | | | | 8.5% | 66% | 87% |
| | PC17 | Häufigkeit Gefühl Niedergeschlagenheit, Angst und Depression (0-10) | Erstellung logistische Regression mit allen verfügbaren Fällen in den Jahren 2021 und 2022 und Anwendung auf Werte aus dem Jahr 2018. | | | |
| | PC18 | Häufigkeit voller Kraft und Energie (0-10) | $P(\widehat{PC206} = 1)_i = \frac{1}{1 + e^{-(-2.73 + PC17_i \cdot 0.49 + PC18_i \cdot -0.22)}}$ | | | |
| | PC206 | Diagnose depressive Störung (nur 2021 und 2022 abgefragt; binäre Klassifizierung) | *M* = .11<br>*SD* = .16<br>Prävalenzbasierter Threshold bei > .19 | | | |
| P-D | | | | 30.4% | 78% | 87% |
| | PHQ9Score | Validiertes Instrument zur Erfassung der depressiven Symptomatik (0-27) | PHQ-9-Summenwert<br>Moderate bis schwere Depression:<br>PHQ-9-Summenwert $\geq$ 10 | | | |

*Anmerkung*. Die Entscheidung beim Datensatz PSYCHE-D mit dem PHQ-9-Summenwert wurde aus Literatur abgeleitet (Levis, Benedetti & Thombs, 2019; Manea, Gilbody & McMillan, 2012; Moriarty, Gilbody, McMillan & Manea, 2015). Die Schätzung für Sensitivität und Spezifität beim Datensatz PSYCHE-D wurde Moriarty et al. (2015) entnommen. Die Abkürzung SHP steht für den Schweizer Haushalt-Panel und P-D für PSYCHE-D. Source: Swiss Household Panel (SHP).

**3.1.1. Schweizer Haushalt-Panel.** Diese Studie wurde mit den Daten des Schweizer Haushaltspanels (SHP) durchgeführt, das am Schweizerischen Kompetenzzentrum für Sozialwissenschaften FORS angesiedelt ist. Das Projekt wird vom Schweizerischen Nationalfonds unterstützt (SHP Group, 2024). Der Schweizer Haushalt-Panel ist eine in der Schweiz einzigartige sozialwissenschaftliche Längsschnittstudie, wobei jährlich seit 1999 teilweise dieselbe Zufallsstichprobe befragt wird (Tillmann et al., 2022). Zusätzlich wurden in den Jahren 2004, 2013 und 2020 Stichproben nachgezogen. Es werden nur Personen befragt, die in privaten Haushalten leben (Voorpostel et al., 2024). Dabei wird darauf geachtet, dass die Stichprobe repräsentativ für die sieben Schweizer Regionen ist, welche bei der Systematik der Gebietseinheiten für die Statistik (NUTS) definiert werden. Personen werden ausgeschlossen, wenn sie fünf Wellen lang nicht erreicht werden konnten und wenn sie nicht mehr in privaten Haushalten leben. So wurden in den in dieser Arbeit behandelten Jahren 2016 10 028, 2017 9 478, 2018 9 349, 2021 13 098 und 2022 11 469 Personen befragt aus 2016 6 261, 2017 5 954, 2018 5 933, 2021 8 299 und 2022 7 258 verschiedenen Haushalten. Dabei werden Personen aus einem Haushalt einerseits einzeln befragt, andererseits gibt es auch Fragen zum Haushalt, die nur einmal pro Haushalt beantwortet werden. Bei den Fragen zum Haushalt werden Variablen aus den Bereichen Zusammenstellung des Haushalts, Aspekte der Unterkunft, Lebensstandard, finanzielle Situation und Organisation im Haushalt und der Familie abgefragt (Voorpostel et al., 2024). Die Fragen aus dem individuellen Fragebogen lassen sich in 20 Kategorien unterteilen. Für diese Arbeit wurden neben dem Haushaltsfragebogen Variablen aus den Kategorien Arbeit, Freizeit, Gesundheit, Netzwerk, Lebensereignisse und Familie betrachtet, da sich diese empirisch und theoretisch hergeleitet für die Prognose depressiver Störungen eignen. Eine Übersicht der Variablen aus diesen Kategorien ist in Anhang C zu finden. Im Schweizer Haushalt-Panel werden auch psychologische Konstrukte, wie beispielsweise die Kontrollüberzeugung, abgefragt, jedoch nur alle drei Jahre, weswegen diese Variablen nicht für die Untersuchung genutzt werden

konnten. Die Befragung findet jeweils von September bis Februar des nächsten Jahres statt (Voorpostel et al., 2024).

Zur Beantwortung der Fragestellung wurden die Messungen aus den Jahren 2016, 2017 und 2018 verwendet. Dabei wurde betrachtet, ob die erhöhte depressive Symptomatik im Jahr 2018 mit den Prädiktionsvariablen in den Jahren 2016 und 2017 vorhergesagt werden kann. Dies wurde gewählt, da gegen Ende des Jahres 2019 und Anfang 2020 im entsprechenden Befragungszeitraum des Jahres 2019 die Covid-19 Pandemie aufkam, die in der Folge einen Einfluss auf diverse Variablen (insbesondere psychologische) hatte, was zu einer starken Konfundierung der Ergebnisse und zu Problemen der Übertragbarkeit der Befunde hätte führen können (Pfefferbaum & North, 2020; Vindegaard & Benros, 2020).

Da im Schweizer Haushalt-Panel die Diagnose Depression nur 2021 und 2022 abgefragt wurde und kein klinisches Instrument zur Erfassung depressiver Symptome verwendet wird, musste eine abhängige Variable für das Jahr 2018 definiert werden. Wichtige, jährlich abgefragte Variablen für die Erfassung von depressiver Symptomatik im Schweizer Haushalt-Panel sind die Variablen PC17 und PC18. Die Variable PC17 erfasst die Häufigkeit von Niedergeschlagenheit, Hoffnungslosigkeit, Angst und Depression («Wie häufig haben Sie negative Gefühle wie Niedergeschlagenheit, Hoffnungslosigkeit, Angst oder Depressionen? 0 bedeutet "nie" und 10 "immer"»; FORS, 2023). Die Variable PC18 erfasst die Häufigkeit von Energie und Optimismus («Wie häufig sind Sie voll Kraft, Energie und Optimismus? 0 bedeutet "nie" und 10 "immer"»; FORS, 2023). Die Variablen PC17 und PC18 wurden auf Basis der Arbeit von Watson, Clark und Tellegen (1988) gebildet, in welcher eine Skala für den positiven und negativen Affekt entwickelt wurde (PANAS). Dabei wurden die Variablen im Schweizer Haushalt-Panel nicht wortwörtlich entnommen. Die beiden Variablen PC17 und PC18 decken wichtige Aspekte der Definition einer depressiven Störung gemäss ICD-11 ab. Darunter sind depressive Stimmung, Niedergeschlagenheit, Hoffnungslosigkeit und verminderte Energie (WHO, 2024). Auch in der Arbeit von Watson et al. (1988) wurde auf

den Zusammenhang von negativem Affekt und Depression, gemessen mit dem Beck Depression Inventory (BDI), eingegangen. Dabei konnte eine signifikante Korrelation von $r$ = .58 gefunden werden. Auch bezüglich positiven Affektes konnte ein signifikanter Zusammenhang von $r$ = -.36 gefunden werden. Auch eine aktuellere Metaanalyse konnte den Zusammenhang zwischen Depressionen und negativem sowie positivem Affekt belegen (Boumparis, Karyotaki, Kleiboer, Hofmann & Cuijpers, 2016). Darüber hinaus zeigen die beiden Variablen PC17 und PC18 im Schweizer Haushalt-Panel einen Anstieg über die letzten 20 Jahre, was der gesellschaftlichen Zunahme der depressiven Symptomatik in der Schweiz entspricht (Bundesamt für Statistik, 2022; Anhang D).

Mit diesen beiden Variablen als unabhängige Variablen (PC17 und PC18) und der Variable PC206 zur Diagnose Depression («Haben Sie eines von den folgenden gesundheitlichen Problemen gehabt, sei es seit (Monat/Jahr) oder chronisch? Damit meinen wir, dass ein Arzt oder eine Ärztin Ihnen das diagnostiziert hat. Depression.»; FORS, 2023) als abhängige Variable wurde eine logistische Regression mit allen verfügbaren Fällen aus den Jahren 2021 und 2022 gerechnet (9 959 Fälle mit einer Prävalenz von 10.8%). Dies bietet die Möglichkeit, aus beiden relevanten Variablen PC17 und PC18 einen nach Prognosefähigkeit gewichteten Wert zu bilden. Beide Variablen haben einen signifikanten Effekt bei $\alpha$ = .05, wobei der Effekt von PC17 positiv ($b$ = 0.49, $p$ <.001) und der von PC18 negativ ($b$ = -0.22, $p$ <.001) ist. Da die Interaktion der beiden Variablen nicht signifikant ist, wurde sie nicht in das Modell einbezogen. Um die abhängige Variable zu bilden, wurde das erstellte Modell in einem ersten Schritt auf die Fälle in den Jahren 2021 und 2022 angewendet. Diese führte zu einem mittleren Wert von $M$ = .11 mit einer Standardabweichung von $SD$ = .16 über alle 9 959 Fälle hinweg.

Weiterführend wurde die erstellte logistische Regression auf die PC17 und PC18 Werte im Jahr 2018 angewendet, da dieser Zeitpunkt zur Definition der abhängigen Variable verwendet wurde. Eine Möglichkeit der Wahl des Thresholds und somit der Zuteilung wäre

bei normalverteilten Daten die Zuteilung über die Standardabweichung. Dabei entspricht eine Standardabweichung bei normalverteilten Variablen über dem Durchschnitt ca. 15.9% der Werte (Moore, McCabe & Craig, 2014). Da die Verteilung der Schätzungen im Jahr 2018, welches zur Klassifizierung der Fälle diente, nicht normalverteilt, sondern linkssteil war, konnte der Threshold nicht entsprechend gewählt werden. Folglich wurde ein Threshold gesucht, welcher zu einer ähnlichen Verteilung der Werte führte, wie in der Messung der Diagnose Depression mit PC206 in den Jahren 2021 und 2022. Dies entsprach einem Anteil von 11.8% vor dem Entfernen der Fälle mit fehlenden Werten im Jahr 2018, was etwa der ursprünglichen Prävalenz entspricht und wurde bei einer halben Standardabweichung über dem Durchschnitt gefunden ($Threshold_{\text{Erhöhte depressive Symptomatik}} > .19$; prävalenzbasierter Threshold als pragmatische Operationalisierung; Balayla, 2024). Dies führte, angewendet auf die vorliegenden Werte aus den Jahren 2021 und 2022 mit der Diagnose Depression (PC206), zu einer Sensitivität von 66.3% und einer Spezifität von 87.2%, wobei in Anhang D zu erkennen ist, dass die mittlere Ausprägung der Symptomatik Unterschiede aufweist zwischen den Jahren 2018 und 2021 sowie 2022 und somit Sensitivität und Spezifität vorsichtig interpretiert werden müssen in Anbetracht einer Diagnose. Dies zeigt den Kompromiss, der bei unausgewogenen Daten eingegangen werden muss: Einerseits braucht es genügend Fälle beider Klassen für die Machine Learning Verfahren, andererseits auch einen klaren und trennscharfen Threshold (He & Garcia, 2009). Ein Threshold, welcher Youden (Summe aus Sensitivität und Spezifität minus eins) optimiert, hätte zu einer überproportional hohen Anzahl an positiven Fällen geführt, wobei ein Grossteil von diesen in Wahrheit keine Diagnose nach PC206 diagnostiziert gehabt hätten (tiefe Präzision; Youden, 1950). So steht in dieser Untersuchung weniger das Erfüllen einer Diagnose, sondern mehr das Erleben von depressiver Symptomatik in einem kritischen Bereich und somit Leidensdruck der Betroffenen im Vordergrund, was hier gewichtet anhand der Diagnose Depression mit den Variablen PC17 und PC18 erreicht werden konnte. Dies ist weiterführend relevant, da auch

soziodemographische Aspekte einen Einfluss darauf haben, ob es zu einer Diagnose kommt und nicht nur die vorhandenen Symptome (Williams et al., 2017).

Nachdem die Fälle über die Operationalisierung der abhängigen Variable den Kategorien zugeteilt wurden, wurde in einem nächsten Schritt der Datensatz der Haushalte mit dem individuellen Datensatz der Personen über das Befragungsjahr und die Haushaltskennung fusioniert. So standen anfangs 1 253 Variablen zur Verfügung. In einem nächsten Schritt wurden die Fälle aus den Jahren 2016 und 2017 ausgewählt und alle weiteren Zeilen ausgeschlossen, da die Prognosen nur in diesen Jahren gemacht wurden, um möglichst aktuelle, aber nicht von Covid-19 verfälschte Ergebnisse zu erhalten. Da das Ziel dieser Arbeit ist, unter anderem mit Hilfe von arbeitsbezogenen Variablen die Vorhersage von depressiven Symptomen durchzuführen, wurden nur Personen ausgewählt, die in beiden Prognosejahren arbeitstätig waren. Darauffolgend wurden nur Variablen aus den beschriebenen interessierenden und theoretisch sowie empirisch hergeleiteten Kategorien ausgewählt (im Schweizer Haushalt-Panel haben diese die Vorzeichen PL, PF, PN, H, PA, PC, PW), was zu 854 verbleibenden Variablen führte. Dabei wird angenommen, dass die Variablen zum Haushalt (H) unter anderem Informationen zum Lebensstil und sozialökonomischen Status beinhalten (Letourneau, Duffet-Leger, Levac, Watson & Young-Morris, 2013; WHO, 2017). In einem nächsten Schritt wurden alle Variablen entfernt, welche mehr als 55% fehlende Werte hatten. Dies wurde einerseits gemacht, um Variablen zu identifizieren, welche nur in einem der beiden Jahre abgefragt wurden (da das Ziel der zeitlich sequentiellen Auswahl die Prognose mit denselben Variablen zu unterschiedlichen Messzeitpunkten ist) und um sicherzustellen, dass es sich nicht um zu spezifische Folgefragen handelt, die nur auf wenige Personen zutreffen. Alle verbleibenden Variablen sind im Anhang C aufgeführt. Folgend wurden gemäss Anhang C die Variablen verarbeitet und einzelne, welche sich nicht zur Prognose eigneten oder in einer anderen Form schon vertreten waren, ausgeschlossen. Likert Skalen mit fünf bis zehn Antwortoptionen wurden als intervallskaliert

betrachtet (Allen & Seamen, 2007; Norman, 2010). Dies führte zu 117 Variablen, welche der automatisierten Variablenauswahl mit Boruta übergeben wurden. Der Datensatz wurde in zwei Datensätze unterteilt, einer für das Jahr 2016 und ein weiterer für das Jahr 2017 und Fälle mit fehlenden Werten wurden entfernt (was im Datensatz 2016 zu 1 709 Fällen führte und 2017 zu 1 701).

Daraufhin wurde mittels OOB-Kurve die Anzahl Bäume für die Variablenselektion auf 200 festgelegt, wie im Anhang E zu erkennen ist. Boruta wurde auf beide Datensätze angewendet und um Variablen auszuwählen, welche einen konstanten Einfluss haben und um aus im Theorieteil beschriebenen Gründen die Anzahl Variablen gering zu halten, wurden nur die Variablen ausgewählt, welche in beiden Datensätzen von Boruta als relevant für die Prognose klassifiziert wurden. Daraufhin wurde aus dem ursprünglichen Datensatz ein neuer Datensatz erstellt, welcher nur die ausgewählten Variablen, die Personenkennung, die abhängige Variable sowie das Geschlecht und Alter zum ersten Messzeitpunkt 2016 und das Jahr der Messung beinhaltete. Erneut wurden fehlende Werte entfernt und nur Personen ausgewählt, bei welchen zu beiden Messzeitpunkten 2016 und 2017 alle Messungen vorlagen. Dies führte zum finalen Datensatz mit $N = 3\ 371$ Personen mit jeweils zwei Messzeitpunkten von $P = 13$ Variablen, welche zur Prädiktion genutzt wurden. Zur Bestätigung der ausgewählten Variablen wurde erneut Boruta angewendet und alle Variablen wurden erneut als relevant klassifiziert. Auf Grund von systematischen Stichprobenausfällen verringerte sich der Anteil an Personen, welche 2018 erhöhte depressive Symptomatik vorwiesen, auf 8.5%. Das durchschnittliche Alter in der finalen Stichprobe ist $M = 44.2$ Jahre, der Median $Mdn = 46.0$ Jahre bei einer Spannweite von 14 bis 66 Jahren. 51.3% der Stichprobe haben männliches Geschlecht angegeben und 48.7% weibliches Geschlecht.

Mit denselben Personen wurde noch pro Prognosezeitpunkt ein weiterer Datensatz erstellt, welcher neben Alter, Geschlecht, Messzeitpunkt und der abhängigen Variable nur noch eine weitere Variable erhielt. Diese Variable enthielt die aktuelle depressive

Symptomatik zu den Prognosezeitpunkten. Dies wurde berechnet mit dem erstellten logistischen Regressionsmodell (erstellt mit Daten aus den Jahren 2021 und 2022 mit PC206 als abhängiger Variable), welches auf die Messungen der Variablen PC17 und PC18 in den Jahren 2016 und 2017 angewendet wurde. Daraufhin wurde die Quantilfunktion einer Standardnormalverteilung (Mittelwert von 0, Standardabweichung von 1) angewendet, um die Voraussetzung der Normalverteilung für die spätere Diskriminanzanalyse zu erfüllen (Buchan, 2025; Conover, 1999; Ghojogh & Crowley, 2019; R Core Team, 2021). Dies wurde gemacht, um einen Vergleich zu haben, ob die ausgewählten zusätzlichen Variablen und Random Forest Modelle die Prädiktionsleistung verbessern oder ob bereits die Messung der depressiven Symptomatik zum Prognosezeitpunkt an sich ausreichend ist für die Prognose, ob die zukünftige depressive Symptomatik in einem kritischen Bereich liegt. Es wurde eine Korrelation des berechneten Symptome-Wertes von $r = .63$ zwischen den Messzeitpunkten 2016 und 2018 und von $r = .67$ zwischen den Messzeitpunkten 2017 und 2018 gefunden.

**3.1.2. PSYCHE-D.** Die verwendeten Daten stammen aus dem PSYCHE-D Datensatz von Makhmutova, Kainkaryam, Ferreira, Min, Jaggi und Clay (2021a), lizenziert unter Cretaive Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). Der Datensatz PSYCHE-D wurde im Rahmen des DiSCover-Projektes von Januar 2018 bis Januar 2019 in den USA erhoben und ist einer der grössten längsschnittlichen Datensätze, der die Ausprägung der depressiven Symptome erfasst (Makhmutova et al., 2021b). Der Datensatz PSYCHE-D ist ein bereits weiterverarbeiteter Datensatz der Rohdaten aus dem DiSCover-Projekt (Makhmutova et al., 2021b). Insgesamt wurden 10 036 Personen befragt. Während der ganzen Studie waren die Teilnehmenden gebeten, ein Wearable zu tragen, welches die Schritt- und Schlafdaten aufzeichnete. Vor Beginn der Studie hatten die Teilnehmenden einen Fragebogen ausgefüllt, welcher Komorbiditäten und soziodemografische Aspekte erfasste. Einmal monatlich waren die Teilnehmenden zusätzlich gebeten, einen Fragebogen zu Veränderungen des Lebensstils und der Medikation und alle drei Monate den PHQ-9

Fragebogen zu Erfassung der depressiven Symptomatik auszufüllen. Die Daten aus den Wearables zu Schritten und zum Schlaf wurden auf drei verschiedene Arten für den Datensatz PSYCHE-D verarbeitet. Erstens wurden statistische Trends über Zeitfenster von vier, sieben und vierzehn Tagen aggregiert. Dabei war immer der Zeitpunkt der PHQ-9 Befragung der Ausgangswert und die Daten wurden rückwirkend aggregiert für die Kennzahlen Mittelwert, Median, Interquartilsabstand und Spannweite. Zweitens wurden Änderungen der Gewohnheiten in den letzten sieben und vierzehn Tagen mittels linearer Regression erfasst. Dabei wurden die Koeffizienten, Intercepts und Regressionsfehler als Variablen abgeleitet. Drittens wurden schwellenbasierte Merkmale verwendet, welche die Anzahl der Tage in den vergangenen sieben und vierzehn Tagen erfassten, in denen ein definierter Schwellenwert über- oder unterschritten wurde. Dabei wurde als Hypersomnie-Tag definiert, wenn mehr als zehn Stunden geschlafen wurde und als Hyposomnie-Tag, wenn weniger als 5 Stunden geschlafen wurde. Aktive Tage wurden mit mehr als 10 000 Schritten definiert und sedentäre Tage mit weniger als 5 000 Schritten (Makhmutova et al., 2021a).

Der in dieser Arbeit verwendete Datensatz PSYCHE-D wurde bereits bereinigt. So wurden für das DiSCover-Projekt ursprünglich 25 508 Personen gescreent, von denen 10 036 Personen in die Untersuchung aufgenommen und befragt wurden. Diese wurden durch Entfernen von Teilnehmenden, die die Qualitätskontrolle nicht bestanden hatten, auf 9 961 reduziert. Danach wurde geprüft, ob die Teilnehmenden mindestens zwei aufeinanderfolgende PHQ-9 Fragebogen ausgefüllt und die Antworten zum Lebensstil am Ende des Quartals beantwortet haben, was die Anzahl weiter reduzierte auf 8 582 Personen. Zuletzt wurde noch geprüft, ob genügend biometrische Daten vorlagen. Dafür wurde jeweils in den zwei Wochen vor Beantwortung der Fragebogen geprüft, an wie vielen Tagen die Wearables mindestens zehn Stunden getragen wurden. Wurden die Wearables nicht mindestens an vier Tagen mehr als zehn Stunden getragen, wurde der Messwert entfernt und die Person ausgeschlossen, was

zu den finalen, im ursprünglichen Datensatz vorhandenen 4 036 Personen führt (Makhmutova et al., 2021b). Dabei liegen 35 694 Messungen vor von 157 Variablen.

Da die Anzahl an validen Messungen in den Monaten, in denen auch der PHQ-9 vorgelegt wurde, im Verlauf des Jahres abnahm (Startmonat 3 676, Monat drei 3 244, Monat sechs 2 832, Monat neun 2 732 und Monat zwölf 2 058), wurden für die vorliegende Arbeit zur Prognose der Monat drei und der Monat sechs verwendet, da für den Startmonat noch keine biometrischen Messungen vorlagen. Dabei wurde für die Definition der abhängigen Variable der Monat neun verwendet. Da für die Messung der depressiven Symptomatik der PHQ-9 verwendet wurde, welcher als valide und reliabel gilt, konnte für die Zuteilung der mehrmals metaanalytisch bestätigte Wert von grösser oder gleich zehn für die Depression verwendet werden, was gemäss PHQ-9 einer moderaten Depression entspricht (Levis, Benedetti & Thombs, 2019; Manea, Gilbody & McMillan, 2012; Moriarty, Gilbody, McMillan & Manea, 2015). Die Sensitivität mit diesem Threshold entspricht gemäss Moriarty et al. (2015) 78% und die Spezifität 87%. Dies führte vor dem Entfernen der fehlenden Werte zu einem anfänglichen Anteil an Personen mit erhöhten depressiven Symptomen von 31.1%.

Nachdem die abhängige Variable definiert wurde, wurde sichergestellt, dass die kategorialen Variablen als kategorial hinterlegt waren und die Metrischen als metrisch. Daraufhin wurden Variablen entfernt, welche mehr als 10% fehlende Werte hatten, um sicherzustellen, dass genügend Messungen für die automatisierte Variablenauswahl mit Boruta vorhanden waren (van der Ploeg et al., 2014). Ausserdem wurden Variablen entfernt, welche eine Korrelation von $r > .7$ (sehr hohe Korrelation), hatten, wobei jeweils die Variable beibehalten wurde, welche mit den anderen Variablen im Durchschnitt eine niedrigere Korrelation aufwies (Kuckartz, Rädiker, Ebert & Schehl, 2013). Dies wurde gemacht, um sicherzustellen, dass genügend Fälle für die Anzahl Variablen zur Verfügung stehen und um verzerrte Schätzungen der Variablenwichtigkeit (und somit die Auswahl durch Boruta) wegen stark korrelierter Variablen zu vermeiden (Barreñada et al., 2024; Rotari & Kulahci, 2024).

Die Gefahr der hohen Korrelationen bestand, da mehrere Variablen aus denselben mit Wearables gemessenen Werten gebildet wurden. Dies führte zu 54 Variablen, welche Boruta zur Variablenselektion übergeben wurden, wobei eine davon die abhängige Variable war. Eine Liste dieser Variablen ist im Anhang F zu finden. Wiederum wurde wie beim Schweizer Haushalt-Panel für jeden Prognosezeitpunkt ein Datensatz gebildet und die Fälle mit fehlenden Werten entfernt, wobei bei Monat drei 1 784 Beobachtungen und bei Monat sechs 1 967 Beobachtungen übrig blieben. Daraufhin wurde mittels OOB-Kurve die Anzahl Bäume für die Variablenselektion auf 200 festgelegt, wie im Anhang E zu erkennen ist. Boruta wurde auf beide Datensätze angewendet und um nur Variablen zu berücksichtigen, welche einen konstanten Einfluss hatten, wurden nur Variablen ausgewählt, welche in beiden Datensätzen als relevant bewertet wurden.

Folgend wurde wiederum ein neuer Datensatz erstellt, welcher neben der abhängigen Variable, der Personenkennung, dem Geschlecht, dem Alter und dem Messzeitpunkt nur die ausgewählten Variablen enthielt. Fälle mit fehlenden Werten wurden entfernt und es wurden nur Personen ausgewählt, welche zu beiden Messzeitpunkten keine fehlenden Werte hatten. Dies führte zum finalen Datensatz mit $N = 1\ 746$ Personen mit jeweils zwei Messzeitpunkten von $P = 16$ Variablen, welche zur Prädiktion genutzt wurden. Der finale Anteil an Personen, die mit erhöhter depressiver Symptomatik im Monat neun klassifiziert wurden, beträgt 30.4%. Das durchschnittliche Alter im finalen Datensatz beträgt $M = 42.2$ Jahre, wobei der Median bei $Mdn = 41$ Jahren liegt. Die Spannweite reicht von 22 bis 82 Jahren. Beim Geschlecht ist der Anteil an Frauen (75.7%) deutlich grösser als der Anteil an Männern (24.3%).

Mit denselben Personen wurde noch pro Prognosezeitpunkt ein weiterer Datensatz erstellt, welcher neben Alter, Geschlecht, Messzeitpunkt und der abhängigen Variable nur noch eine zusätzliche Variable erhielt. Diese enthielt den PHQ-9-Wert zu den Prognosezeitpunkten, welcher mit der van der Waerden Methode einer Normalverteilung angepasst wurde. Dabei wird bei der van der Waerden Methode eine Rangverteilung

durchgeführt, welche durch $N + 1$ geteilt wird, um Werte zwischen null und eins zu erhalten (Buchan, 2025; Conover, 1999). Daraufhin wurde die Quantilfunktion einer Standardnormalverteilung angewendet (Buchan, 2025; Conover, 1999; R Core Team, 2021). Dies wurde gemacht, um eine Vergleichbarkeit zu haben, ob die ausgewählten Variablen die Prädiktionsleistung verbessern oder ob bereits die Messung der depressiven Symptomatik zum Prognosezeitpunkt an sich ausreichend ist für die Prognose, ob die zukünftige depressive Symptomatik in einem kritischen Bereich liegt. Es wurde eine Korrelation des berechneten Symptome-Wertes von $r = .73$ zwischen den Messzeitpunkten Monat drei und Monat neun und von $r = .75$ zwischen den Messzeitpunkten Monat sechs und Monat neun gefunden.

## 3.2. Statistische Modellierung

Im folgenden Teil der Arbeit wird zuerst ein Überblick gegeben über das weitere Vorgehen nach der Erstellung der Datensätze. Darauffolgend werden die Verfahren genauer beschrieben.

**3.2.1. Übersicht des Vorgehens.** Eine übersichtliche Darstellung des Vorgehens ist in Abbildung 1 zu finden. Dabei wurden mehrere Verarbeitungsschritte aufeinander abgestimmt. Es wurde zuerst jeweils anhand der Trainingsdaten (vier der fünf äusseren Folds) pro Prognosezeitpunkt ein Random Forest Modell erstellt, wobei in der inneren Schleife die Hyperparameter gewählt wurden (nested Cross Validation). Die finalen Random Forests mit den gewählten Hyperparametern wurden von allen jeweiligen Trainingsdaten erstellt und auf diese angewendet, was zu einer in-sample Prognose führte. Mit diesen Prognosen und den wahren Werten der künftigen erhöhten depressiven Symptomatik wurden auf Basis der Trainingsdaten die Modelle und Werte für die Diskriminanzanalyse geschätzt. Darauffolgend wurden jeweils mit den durch die Trainingsdaten erstellten Random Forests Prognosen für den verbleibenden äusseren Fold (ein Fünftel der Daten; Testdaten) geschätzt. Diese Schätzungen wurden wiederum den mithilfe der Trainingsdaten geschätzten

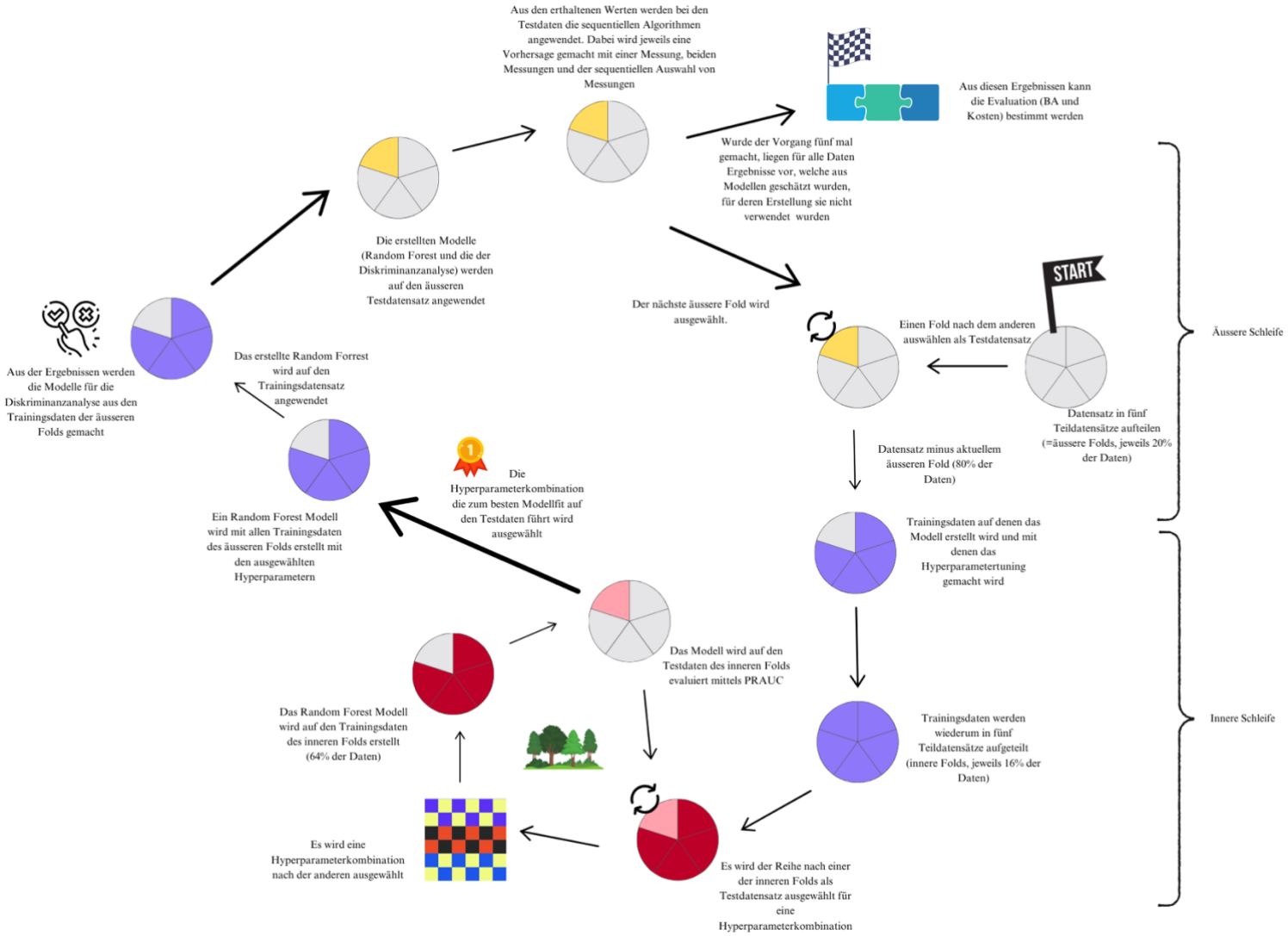


*Abbildung 1.* Vorgehen bis zur Anwendung der sequentiellen Algorithmen. *Anmerkung.* Eigene Abbildung in Anlehnung an Weaverdyck, Lieberman und Parkinson (2020, S. 502)

Diskriminanzmodellen übergeben, was zu den finalen Werten führte, die den Algorithmen zur sequentiellen Auswahl und Prognose mit POSEIDON übergeben wurden (Wyss et al., 2023). Für die Prognose nur mit der aktuellen depressiven Symptomatik wurde nur die Diskriminanzanalyse mit 5-facher Cross-Validation durchgeführt, bevor POSEIDON angewendet wurde.

**3.2.2. Random Forests zur Implementierung der prädiktiven Modelle.** Nach der Datensatzbildung lag für die beiden Datensätze Schweizer Haushalt-Panel und PSYCHE-D für jeden Prognosezeitpunkt ein Datensatz mit denselben Variablen und Fällen vor. In einem ersten Schritt wurden die Datensätze in fünf Teildatensätze/Folds unterteilt (äussere Folds; $k_{outer} = 5$), wobei darauf geachtet wurde, dass Personen in beiden Datensätzen der unterschiedlichen Messzeitpunkte demselben Teildatensatz angehörten (Hastie et al., 2017). Ausserdem wurde sichergestellt, dass in allen Teildatensätzen derselbe Anteil an Personen mit und ohne künftige depressive Symptomatik im kritischen Bereich vertreten waren.

In einem nächsten Schritt wurde mittels nested Cross Validation das Hyperparametertuning pro Messzeitpunkt durchgeführt. Dabei wurden für jeden äusseren Fold die Fälle aus dem Fold als Testdaten verwendet und die Fälle des Datensatzes, welche nicht im Fold waren, als Trainingsdaten. Die Trainingsdaten wurden wiederum für das Hyperparametertuning in fünf innere Folds unterteilt, wobei wiederum darauf geachtet wurde, dass in allen Teildatensätzen derselbe Anteil an Personen der Kategorien der abhängigen Variable waren ($k_{inner} = 5$; Hastie et al., 2017). Folgend wurden für jede Hyperparameterkombination Random Forests anhand von vier der fünf inneren Folds trainiert und am fünften inneren Fold evaluiert, wobei PRAUC als Mass für die Prognosegüte verwendet wurde (Hancock et al., 2023). Die Hyperparameter, welche jeweils in allen möglichen Kombinationen getestet wurden, sind in Tabelle 2 abgebildet. Die Anzahl Bäume wurde wiederum mit der OOB-Kurve auf 200 festgelegt (siehe Anhang E). Dabei gilt zu beachten, dass beim Datensatz PSYCHE-D fälschlicherweise die Variable «Fold», welche die

Tabelle 2
*Verwendete Hyperparameter für das Hyperparametertuning*

| Hyperparameter | Schweizer Haushalt-Panel | PSYCHE-D |
|---|---|---|
| Anzahl Bäume | 200 | 200 |
| Regel für die Entscheidung | Hellinger, Extratrees, Gini | Hellinger, Extratrees, Gini |
| Anzahl zufällig ausgewählte Variablen für jeden Split | 2, 4, 6, 8, 10, 12 | 2, 4, 6, 8, 10, 12, 14, 16 |
| Minimale Anzahl Fälle pro Knoten | 5, 10, 20, 30, 40 | 5, 10, 20, 30, 40 |

Zugehörigkeit zur äusseren Schleife beinhaltet, der Erstellung der Random Forests übergeben wurde, was jedoch einen geringen Einfluss auf die Modellleistung und Schätzungen hat, was in den Limitationen genauer diskutiert wird.

Nachdem die beste Hyperparameterkombination pro Messzeitpunkt mittels der inneren Folds gefunden wurde, wurde in einem nächsten Schritt das Modell am gesamten aktuellen Trainingsdatensatz (vier der fünf äusseren Folds) erstellt und Prädiktionen mit dem erstellten Modell für die Trainingsdaten gemacht (vier der fünf äusseren Folds; jeweils separat für die beiden Datensätze des Schweizer Haushalt-Panels und PSYCHE-D sowie jeweils für die beiden Messzeitpunkte). Anschliessend wurde das erstellte Random Forest Modell auf den verbleibenden äusseren Fold, sprich die jeweiligen Testdaten, angewendet.

Zur Erfassung der Variablenwichtigkeit wurde die Mean Decrease in Impurity (MDI) Methode verwendet. Diese betrachtet bei Klassifizierungen, inwiefern Variablen, wenn sie bei einem Split ausgewählt wurden, die Unreinheit verringern (Breiman, 2001). Diese wurde über beide Messzeitpunkte und alle äusseren Folds hinweg gemittelt berechnet bei den finalen Random Forest Modellen. Dabei wird zur Vergleichbarkeit jeweils pro Fold die wichtigste Variable auf *MDI* = 100.0 gesetzt und die weiteren Prädiktoren werden relativ dazu eingeschätzt.

**3.2.3. Regressionsbasierte Diskriminanzmodelle.** Die Vorhersagen basierend auf den Random Forests wurden weiterverarbeitet, um sie mit der Annahme der Normalverteilung, welche eine Voraussetzung für die Diskriminanzanalyse darstellt, in Einklang zu bringen. Die Transformation wurde mit der Funktion «qnorm()» aus dem R Base-Paket «stats» durchgeführt (Quantilfunktion einer Standardnormalverteilung; R Core Team, 2021). Daraufhin wurden wiederum mit den Trainingsdaten die Modelle für die Diskriminanzanalyse erstellt (siehe Abbildung 1). Somit unterscheiden sich die Modelle je nach Fold *k*. Die Modelle für die Diskriminanzanalyse wurden mit in-sample Vorhersagen der Trainingsdaten der Random Forests erstellt. Dabei wurde ein Prävalenzmodell für beide Messzeitpunkte, ein multivariates Markermodell (Vorhersagen der Mittelwerte für beide Klassen und Kovarianzmatrix der Residuen) für beide Messzeitpunkte berechnet. Der Wert *z* entspricht der Klassenzugehörigkeit bei der operationalisierten abhängigen Variable und zeigt somit erhöhte künftige depressive Symptomatik. Für das Vorgehen ist es wichtig, dass der wahre Wert *z* der Bernoulli Verteilung folgt und die Markerwerte (Prognosewerte aus dem Random Forests) pro Zeitpunkt $\boldsymbol{y}_i = \begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix}$ ($y_{i,t}$ Markerwert zum Zeitpunkt $t$) gegeben der Klasse *z* normalverteilt ist, was in Gleichungen 1 dargestellt ist (adaptiert von Wyss et al., 2023).

$$z_i \sim Bernouli\left(\pi_{1,0}\right)$$
$$\boldsymbol{y}_i | z \sim N_2\left(\boldsymbol{\mu}_i^{(z)}, \boldsymbol{\Sigma}\right) \; (z \in \{1; 2\}) \tag{1}$$

Der Vektor $\boldsymbol{\mu}_i^{(z)} = \begin{pmatrix} \mu_{i,1}^{(z)} \\ \mu_{i,2}^{(z)} \end{pmatrix}$ enthält die klassenspezifischen Mittelwerte beim ersten ($E\left(y_{i,1}|z\right) = \mu_{i,1}^{(z)}$) und zweiten ($E\left(y_{2,1}|z\right) = \mu_{i,1}^{(z)}$) Messzeitpunkt und die Matrize $\boldsymbol{\Sigma}_i =$

$\begin{pmatrix} Var(y_{i,1}) & Cov(y_{i,1}, y_{i,2}) \\ Cov(y_{i,1}, y_{i,2}) & Var(y_{i,2}) \end{pmatrix}$ die Varianzen beim ersten und zweiten Messzeitpunkt sowie die Kovarianz zwischen den beiden Messzeitpunkten. Das Prävalenzmodell wurde bei den Trainingsdaten anhand der wahren Klassenzugehörigkeit bei der operationalisierten abhängigen Variable *z* berechnet über eine logistische Regression. Gleichung 2 zeigt, wie die Prävalenz $\pi_{0,i}$ einer Person $i$ anhand einer logistischen Regression mit dem Alter $x_{i,1}$ und dem Geschlecht $x_{i,2}$ als Prädiktoren (Prävalenzmodell) vorhergesagt wird.

$$\pi_{0,i} = P(z = 1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2)}} \tag{2}$$

Mithilfe des Trainingsdatensatzes (jeweils vier der fünf äusseren Folds) wurden die Effekte $\hat{b}_0$, $\hat{b}_1$ und $\hat{b}_2$ geschätzt, so dass die Prävalenz für Personen ausserhalb des Trainingsdatensatzes vorhergesagt wurde ($\hat{\pi}_{0,i}$). Somit kann anhand der Verteilung in den Trainingsdaten mit den beiden Variablen Alter und Geschlecht bei den Testdaten eine Prävalenz geschätzt werden, welche der a-priori Wahrscheinlichkeit entspricht (Hastie et al., 2017; WHO, 2017).

Für das Markermodell wurde eine multivariate lineare Regression mit den einer Standartnormalverteilung angepassten Markerwerten der Random Forest beider Zeitpunkte verwendet, in welchem die Mittelwerte modelliert wurden in Abhängigkeit der Klasse $z_i$, dem Alter $x_{i,1}$ und dem Geschlecht $x_{i,2}$ sowie der Interaktionen aus der Klasse $z_i$ und dem Alter $x_{i,1}$ sowie der Klasse $z_i$ und dem Geschlecht $x_{i,2}$. In der multivariaten Regression wird angenommen, dass

$$y_{i,t} = b_{t,0} + b_{t,1}x_{1,i} + b_{t,2}x_{2,i} + b_{t,3}z_i + b_{t,4}x_{1,i}z_i + b_{t,5}x_{2,i}z_i + \varepsilon_{i,t} \; (t \in \{1; 2\})$$

$\boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \end{pmatrix} \sim N_2(0, \boldsymbol{\Sigma})$, $\boldsymbol{\varepsilon}_i$ und $\boldsymbol{\varepsilon}_j$ unabhängig für $i \neq j$. Alle Modelparameter $b_{1,0}, \ldots, b_{1,5}$, $b_{2,0}, \ldots, b_{2,5}$ und $\boldsymbol{\Sigma}$ wurden mit Trainingsdaten geschätzt, so dass für Personen ausserhalb des Trainingdatensatzes die Verteilungen von $\boldsymbol{y}_i|z = 1$ und $\boldsymbol{y}_i|z = 2$ vorhergesagt werden konnte pro Messzeitpunkt. Es wurde angenommen, dass $\boldsymbol{y}_i|z$ einer Normalverteilung folgt mit Mittelwert

$$\hat{\mu}_{i,t}^{(z)} = \hat{b}_{t,0} + \hat{b}_{t,1} x_{1,i} + \hat{b}_{t,2} x_{2,i} + \hat{b}_{t,3} z + \hat{b}_{t,4} x_{1,i} z_i + \hat{b}_{t,5} x_{2,i} z \; (z \in \{1; 2\}, t \in \{1; 2\})$$

und Kovarianzmatrix $\widehat{\boldsymbol{\Sigma}}$ (geschätzte Kovarianzmatrix der Residuen $\hat{\boldsymbol{\varepsilon}}_i = \begin{pmatrix} y_{i,1} - \hat{\mu}_{i,1}^{(z)} \\ y_{i,2} - \hat{\mu}_{i,2}^{(z)} \end{pmatrix}$).

Die Modelle und Kovarianzmatrixen wurden jeweils anhand der Trainingsdaten und in-sample Vorhersagen der Random Forests geschätzt und auf die Testdaten angewendet.

**3.2.4. Sequentielle Auswahl von Messzeitpunkten.** Auf Basis der Schätzungen aus den Random Forests und den Modellen sowie Werten aus der Diskriminanzanalyse wurden Prognosen gemacht für die spätere erhöhte depressive Symptomatik. Dabei wurden Klassifizierung gemacht, welche jeweils nur die erste Messung, jeweils beide Messungen oder eine sequentielle Auswahl an Messungen (basierend auf POSEIDON) verwendeten. Es wurden jeweils mit den erstellten Modellen verschiedene Prognosen gemacht, wobei die Kosten der Fehlklassifikationen für falsch positiv und falsch negativ und die Kosten der Messung variiert wurden. Die Kosten für die Fehlklassifikation wurden entweder gleich oder invers proportional zum Anteil der wahren Gruppe gewichtet. Gegeben den Kosten für Fehlklassifikationen (falsch positiv und falsch negativ) ergeben sich unterschiedliche Cutoffs für die Klassifikation anhand der posterior Wahrscheinlichkeit (Elkan, 2001): Bei gleicher Gewichtung Cutoff = .5 und bei Gewichtung nach inversem Anteil Cutoff = Anteil Gruppe erhöhte depressive Symptomatik. Bei gleicher Gewichtung der Fehlklassifikationskosten wurde für falsch positiv und falsch negativ jeweils der Wert eins gewählt. Bei der Gewichtung

nach inversem Anteil wurden die Kosten für falsch positive Klassifizierungen wie in Gleichung 3 berechnet.

$$Kosten_{Falsch\ Negativ} = \frac{1}{2\ \cdot\ Anteil_{Erhöhte\ depressive\ Symptomatik}} \tag{3}$$

Dies entspricht beim Datensatz Schweizer Haushalt-Panel falsch negativen Fehklassifikationskosten von $1/(0.085 \cdot 2) = 5.88$ und bei PSYCHE-D $1/(0.304 \cdot 2) = 1.64$. Die falsch positiven Messkosten bei der Gewichtung nach inversem Anteil wurden wie in Gleichung 4 zu erkennen ist berechnet.

$$Kosten_{Falsch\ Positiv} = \frac{1}{2\ \cdot\ (1 - Anteil_{Erhöhte\ depressive\ Symptomatik})} \tag{4}$$

Dies entspricht beim Datensatz Schweizer Haushalt-Panel $1/(0.915 \cdot 2) = 0.55$ und beim Datensatz PSYCHE-D $1/(0.696 \cdot 2) = 0.72$. Dies entspricht bei der ungleichen Gewichtung dem inversen Anteil der Klasse und ist somit umgekehrt proportional zur Grösse dieser (Araf, Idri & Chairi, 2024; Razzaghi, Roderick, Safro & Marko, 2016). Es wurde durch zwei geteilt, da die BA verwendet wurde zur Berechnung der Prognosegüte, welche über den Mittelwert von Sensitivität und Spezifität berechnet wird (Broderson et al., 2010).

Die Messkosten wurden folgend jeweils so gewählt, dass als obere Begrenzung so hohe Kosten gewählt wurden, dass bei den sequentiellen Selektionen immer nur eine Messung ausgewählt wurde über alle Personen hinweg. Die untere Begrenzung wurde so gewählt, dass die Prognosegüte, gemessen mit der BA, der Prognose mit zwei starren Messungen durch die sequentielle Auswahl annähernd erreicht wurde. Zwischen diesen Extremwerten wurden die Messkosten im regelmässigen Abstand gewählt. Die Veränderung der Messkosten hat bei den Prädiktionen mit einer festen Auswahl an Messzeitpunkten (einem oder zwei) keinen Einfluss

auf die Prognosegüte, jedoch auf die durchschnittlichen Gesamtkosten. Diese wurden über die Summe aus Anzahl an Messungen multipliziert mit den Kosten einer Messung, Fehlklassifikationskosten für falsch positive Fälle multipliziert mit der Anzahl falsch positiver Zuteilungen und Fehlklassifikationskosten für falsch negative Fälle multipliziert mit der Anzahl falsch negativer Zuteilungen berechnet. Die erhaltene Summe wurde durch die Anzahl Personen geteilt, um die durchschnittlichen Gesamtkosten pro Person zu ermitteln. Sowohl die Anpassung der Kosten für Fehlklassifikationen sowie für Kosten der Messung haben einen Einfluss auf die sequentielle Selektion, da eine unterschiedliche Anzahl an zweiten Messzeitpunkten ausgewählt wird.

Auf die Entscheidung zur Auswahl einer zweiten Messung wird in der Arbeit von Wyss et al. (2023) und in der Dokumentation zu POSEIDON (Universität Bern, 2022) genauer eingegangen. Dabei wird der zweite Messzeitpunkt nur abgewartet, wenn die erwarteten Kosten durch die zweite Messung reduziert werden können. Die Entscheidung zur Auswahl einer zweiten Messung basiert auf den Werten und Modellen aus der Diskriminanzanalyse, welche wie in Kapitel 3.2.2. beschrieben berechnet werden und wiederum auf den Schätzungen aus den Random Forests aufbauen, sowie den gewählten Mess- und Fehlklassifikationskosten (Wyss et al., 2023). Dabei werden individuelle Schwellenwerte festgelegt, welche auf Prävalenz, Kosten und der Mahalanobis Distanz basieren (Maesschalk, Jouan-Rimbaud, Massart, 2000; Wyss et al., 2023). Die Zuteilung erfolgt über den Satz von Bayes und der Cutoff für eine Zuteilung richtet sich nach den gewählten Fehlklassifikationskosten (gleiche Gewichtung oder Gewichtung nach inversem Anteil an Personen mit erhöhter depressiver Symptomatik; Hastie et al., 2017; Wyss et al., 2023).

**3.2.5. Software für die Auswertung.** Für die Auswertung wurde die frei zugängliche Statistiksoftware R Studio verwendet. Es wurde die zum Zeitpunkt der Auswertungen aktuelle Version 4.4.2. verwendet (31.10.2024). Eine Auflistung der verwendeten Pakete und deren

Version ist in Anhang G abgebildet. Die R-Codes und Skripte der vorliegenden Arbeit können aus Gründen der Transparenz und Reproduzierbarkeit beim Autor angefragt werden. Die R-Skripte zur Diskriminanzanalyse und POSEIDON wurden von der Betreuungsperson bereitgestellt.

# 4. Ergebnisse

In einem ersten Schritt werden die Ergebnisse der Variablenauswahl und des Hyperparametertunings vorgestellt, um die erstellten Random Forest Modelle einschätzen zu können. Daraufhin werden die in der Diskriminanzanalyse erstellten Modelle berichtet, bevor auf die Ergebnisse der sequentiellen Algorithmen eingegangen wird. Dies wird jeweils nach Datensatz unterteilt.

## 4.1. Prädiktionsmodelle

In diesem Kapitel werden die Ergebnisse der Variablenauswahl durch Boruta und des Hyperparametertunings vorgestellt. Es werden zuerst die Ergebnisse des Datensatzes Schweizer Haushalt-Panel vorgestellt, bevor auf die Ergebnisse des Datensatzes PSYCHE-D eingegangen wird.

**4.1.1. Schweizer Haushalt-Panel.** Die mittels Boruta zu beiden Messzeitpunkten als relevant klassifizierten Variablen sind in Tabelle 3 zu erkennen. Ergänzend ist die mittlere Variablenwichtigkeit (MDI) aus der späteren Anwendung mit nested Cross Validation über alle Folds und Messzeitpunkte hinweg angegeben. Es ist zu erkennen, dass die Variable Häufigkeit Gefühl Niedergeschlagenheit, Angst und Depression (PC17), welche unter anderem zur Konstruktion der abhängigen Variable verwendet wurde, die deutlich höchste Variablenwichtigkeit aufweist ($M$ = 87.8, $SD$ = 25.7). Weiterführend ist zu erkennen, dass auch die Variable Häufigkeit voller Kraft und Energie (PC18), welche ebenfalls zur Konstruktion der abhängigen Variable verwendet wurde, eine relativ betrachtet hohe Variablenwichtigkeit nach MDI aufweist ($M$ = 24.2, $SD$ = 10.8). Ansonsten wurden besonders

Tabelle 3
*Schweizer Haushalt-Panel durch Boruta ausgewählte Variablen*

| Variablen-kennung | Frage | Mittlere Variable Wichtigkeit *(SD)* |
|---|---|---|
| PC17 | Wie häufig haben Sie negative Gefühle wie Niedergeschlagenheit, Hoffnungslosigkeit, Angst oder Depressionen? 0 bedeutet «nie» und 10 «immer». | 87.8*(25.7)* |
| PW77 | Wie viele Stunden pro Woche schaffen Sie normalerweise in Ihrer Haupttätigkeit? | 37.1*(33.4)* |
| PC50 | Wie häufig empfinden Sie normalerweise folgende Gefühle? 0 bedeutet «nie» und 10 «immer»? Sorge. | 27.9*(11.6)* |
| PC18 | Wie häufig sind Sie voll Kraft, Energie und Optimismus? 0 bedeutet «nie» und 10 «immer». | 24.2*(10.8)* |
| PF51 | Wie stark sind Sie nach der Arbeit zu erschöpft, um Sachen zu machen, wo Sie eigentlich gerne machen, wenn 0 «überhaupt nicht» und 10 «äusserst stark» bedeutet? | 24.0*(20.1)* |
| PC49 | Wie häufig empfinden Sie normalerweise folgende Gefühle? 0 bedeutet «nie» und 10 »immer»? Traurigkeit. | 22.4*(11.5)* |
| PA06 | Wie zufrieden sind Sie mit den Aktivitäten, wo Sie in Ihrer Freizeit machen? 0 bedeutet «gar nicht zufrieden» und 10 bedeutet «vollumfänglich zufrieden»? | 21.3*(15.8)* |
| PF50 | Wie stark haltet Sie die Arbeit, mehr als Ihnen lieb ist, von Ihren privaten Aktivitäten und familiären Verpflichtungen ab, wenn 0 «überhaupt nicht» und 10 «äussert stark» bedeutet? | 21.0*(18.0)* |
| PC48 | Wie häufig empfinden Sie normalerweise folgende Gefühle? 0 bedeutet «nie» und 10 «immer»? Ärger. | 20.0*(15.5)* |
| PC47 | Wie häufig empfinden Sie normalerweise folgende Gefühle? 0 bedeutet "nie" und 10 "immer"? Freude. | 17.7*(9.9)* |
| PC44 | Ganz allgemein gefragt - wie zufrieden sind Sie mit Ihrem Leben? 0 bedeutet «gar nicht zufrieden» und 10 «vollumfänglich zufrieden». | 15.3*(10.3)* |
| PC184 | Wie oft haben Sie Nervosität und Stress empfunden? 1 bedeutet «nie», 2 bedeutet «selten», 3 bedeutet «manchmal», 4 bedeutet «häufig», 5 bedeutet «sehr oft». | 14.9*(8.3)* |
| PC05A | Haben Sie in den letzten 4 Wochen Schmerzen, oder gesundheitliche Beschwerden gehabt, die ich jetzt aufzähle, und zwar "überhaupt nicht", "ein bisschen" oder "stark" ? Allgemeine Schwäche, Erschöpfung, einfach keine Energie. | 2.4*(3.8)* |

*Anmerkung.* Fragen wurden aus der Dokumentation kopiert (FORS, 2023).
Die Variablenwichtigkeit wurde mittels MDI berechnet und über alle Folds und beide Zeitpunkte gemittelt. Die wichtigste Variable pro Fold wird jeweils auf *MDI* = 100 gesetzt und die anderen relativ dazu eingeschätzt. In Klammern ist die Standardabweichung (*SD*) über die Folds hinweg. Die hohe *SD* ist teilweise auf die Korrelation der Variablen zurückzuführen (Strobl et al., 2008; Anhang H). Source: Swiss Household Panel (SHP).

Variablen zur Gesundheit zu beiden Messzeitpunkten von Boruta als relevant klassifiziert.

Werden die im Hyperparametertuning ausgewählten Hyperparameterkombinationen in Tabelle 4 betrachtet, ist zu erkennen, dass kein Hyperparameterwert deutlich öfter vorkommt als die anderen, jedoch Tendenzen zu erkennen sind. Bei der Regel für die Entscheidung fällt auf, dass Hellinger und Extratrees jeweils nur zweimal ausgewählt wurden, währenddem Gini öfter ausgewählt wurde. Bei der Anzahl der zufällig ausgewählten Variablen pro Split fällt auf, dass diese hoch gewählt wurden ($> \sqrt{P}$). Die minimale Anzahl Fälle pro Knoten und somit die Tiefe der einzelnen Entscheidungsbäume unterscheidet sich zwischen den Folds, wobei gemäss PRAUC keine gewählte Tiefe deutliche Vorteile vorweisen kann. Auch dieser Wert weicht vom Default Wert von eins ab und der tiefste dem Hyperparametertuning übergebene Wert fünf wurde nur bei einem Fold ausgewählt. Auffallend ist dabei, dass beim zweiten Messzeitpunkt die kleinste gewählte Tiefe von fünf zum geringsten PRAUC-Wert führt. Die PRAUC Werte, welche verwendet wurden, um die beste Hyperparameterkombination in jedem Fold auszuwählen, sind abhängig vom Anteil der positiven Klasse (Ozenne, Subtil & Maucort-Boulch, 2015). Bei beiden Datensätzen des Schweizer Haushalt-Panels ist der Wert über die Folds hinweg mindestens doppelt so gross, als er bei einer zufälligen Schätzung zu erwarten wäre (Ozenne et al., 2015). Dabei fällt auch auf, dass die PRAUC Werte jeweils beim zweiten Messzeitpunkt tendenziell höher sind als beim Ersten, auch wenn diese Unterschiede gering sind.

Tabelle 4
*Pro Fold ausgewählte Hyperparameter und PRAUC Schweizer Haushalt-Panel*

| Messzeitpunkt | Fold | Regel für die Entscheidung | Anzahl zufällig ausgewählte Variablen für jeden Split | Minimale Anzahl Fälle pro Knoten | PRAUC |
|---|---|---|---|---|---|
| 1 | 1 | Hellinger | 8 | 10 | .20 |
| 1 | 2 | Gini | 8 | 30 | .22 |
| 1 | 3 | Gini | 6 | 30 | .25 |
| 1 | 4 | Gini | 10 | 30 | .20 |
| 1 | 5 | Extratrees | 12 | 10 | .20 |
| 2 | 1 | Extratrees | 8 | 5 | .21 |
| 2 | 2 | Gini | 12 | 20 | .26 |
| 2 | 3 | Gini | 12 | 40 | .30 |
| 2 | 4 | Gini | 10 | 30 | .26 |
| 2 | 5 | Hellinger | 10 | 40 | .25 |

*Anmerkung.* Beim PRAUC gilt zu beachten, dass dieser stark vom Anteil der positiven Klasse abhängig ist und somit im Verhältnis zu dieser einzuschätzen ist (Ozenne et al., 2015). Beim Schweizer Haushalt-Panel betrug diese 8.5%. Die PRAUC-Werte wurden über die inneren Folds bestimmt anhand von denen das Modell mit den entsprechenden Hyperparametern evaluiert und ausgewählt wurde.

**4.1.2. PSYCHE-D.** Die mittels Boruta zu beiden Messzeitpunkten als relevant klassifizierten Variablen sind in Tabelle 5 zu erkennen. Ergänzend ist die mittlere Variablenwichtigkeit aus der späteren Anwendung mit nested Cross Validation angegeben. Es ist zu erkennen, dass die Variable mit dem PHQ-9-Summenwert (PHQ9Score), welche zur Konstruktion der abhängigen Variable verwendet wurde, in jedem Fold als wichtigste Variable über MDI bewertet wurde ($M$ = 100.0, $SD$ = 0.0). Weiterführend ist zu erkennen, dass die biometrischen Messungen eine hohe Standardabweichung aufweisen, was auf die Korrelation der Prädiktionsvariablen zurückgeführt werden kann, da diese teilweise aus denselben Werten berechnet wurden (Strobl, Boulesteix, Kneib, Augustin & Zeileis 2008). Es ist zu erkennen, dass die biometrischen Messungen eine höhere Variablenwichtigkeit aufweisen als die Variablen zum Lebensstil. Dabei wurde die Variable zum Beginn einer neuen Medikation (med_start), bei keinem Fold in einem Split ausgewählt ($M$ = 0.0, $SD$ = 0.0).

Tabelle 5
*PSYCHE-D durch Boruta ausgewählte Variablen*

| Variablenkennung | Variablenbildung | Mittlere Variable Wichtigkeit *(SD)* |
|---|---|---|
| PHQ9Score | Der Wert der PHQ-9 Skala (Werte 0-27). | 100.0*(0.0)* |
| sleep_ratio_asleep_in_bed_weekday_mean | Mittlerer Anteil der Schlafenzeit von gesamter Zeit, die im Bett verbracht wurde über die letzten vierzehn Tage an Wochentagen. | 17.2*(20.7)* |
| steps__rolling_6_sum__max__intercept | Intercept der Regression der maximalen Anzahl Schritte innerhalb eines sechsminütigen Fensters an einem Tag über vierzehn Tage hinweg. | 17.0*(22.3)* |
| sleep__main_efficiency__coeff | Regressionskoeffizient für die Effizienz des Schlafes und somit Veränderung des Trends über die letzten vierzehn Tage (Anteil der geschlafenen Zeit von der gesamten Zeit im Bett) | 16.5*(22.3)* |
| sleep_in_bed_weekday_mean | Mittlere Anzahl an Minuten im Bett pro Tag an Wochentagen über die letzten vierzehn Tage hinweg. | 16.4*(22.6)* |
| sleep__main_start_hour_adj__coeff | Regressionskoeffizient für die Haupteinschlafenszeit und somit Veränderung des Trends über die letzten vierzehn Tage. | 15.5*(20.4)* |
| steps_lpa_sum_recent | Gesamtdauer der Minuten, in denen Schritte im LPA (Light Physical Activity) Bereich gemacht wurden, über die letzten vier Tage hinweg. | 15.4*(18.9)* |
| steps__light_activity__sum__intercept | Intercept der Regression der Anzahl Schritte in leichter Aktivität über die letzten vierzehn Tage hinweg. | 15.2*(20.4)* |
| steps_rolling_6_max_recent | Maximale Anzahl Schritte in einem sechsminütigen Fenster über die letzten vier Tage hinweg. | 14.7*(19.9)* |
| sleep__nap_count__intercept | Intercept der Regression der Anzahl Nickerchen über die letzten vierzehn Tage hinweg. | 12.7*(16.7)* |
| steps_mvpa_sum_recent | Gesamtdauer der Minuten, in denen Schritte im MVPA (Moderate to Vigorous Physical Activity) Bereich gemacht wurden, über die letzten vier Tage hinweg. | 12.5*(17.3)* |
| sleep_main_start_hour_adj_median | Median der Haupteinschlafenszeit über die letzten vierzehn Tage. | 9.9*(9.7)* |
| steps__sedentary_day_count_ | Anzahl Tage mit weniger als 5 000 Schritten über sieben Tage hinweg. | 6.7*(4.3)* |
| sleep__hyposomnia_count_ | Anzahl Nächte mit weniger als fünf Stunden Schlaf in den letzten sieben Tagen. | 6.1*(4.8)* |
| med_nonmed_dnu | Nimmt keine Medikamente und hat auch keine nonmedikamentöse Therapie begonnen. | 2.8*(1.4)* |
| med_start | Hat im letzten Monat ein neues Medikament angefangen zu nehmen. | 0.0*(0.0)* |

*Anmerkung.* Die Angaben beziehen sich auf die Dokumentation der Arbeit von Makhmutova et al. (2021a). Die Variablenwichtigkeit wurde mittels MDI berechnet und über alle Folds und beide Zeitpunkte gemittelt. Die wichtigste Variable pro Fold wird jeweils auf *MDI* = 100 gesetzt und die anderen relativ dazu eingeschätzt. In Klammern wurde die Standardabweichung (*SD*) eingetragen über die Folds hinweg. Die hohe *SD* ist teilweise auf die Korrelation der Variablen zurückzuführen (Strobl et al., 2008, Anhang H).

Die ausgewählten Hyperparameter pro Fold und Datensatz sowie die PRAUC-Werte sind in Tabelle 6 eingetragen. Dabei ist wiederum zu erkennen, dass kein Hyperparameterwert deutlich öfter vorkommt als die anderen, jedoch Tendenzen zu erkennen sind. Bei der Regel für die Entscheidung fällt auf, dass Hellinger nur zweimal ausgewählt wurde, währenddem Extratrees und Gini in etwa ausgeglichen sind. Bei der Anzahl der zufällig ausgewählten Variablen pro Split fällt besonders die grosse Spannbreite auf, wobei als Tendenz Werte zwischen sechs und zehn gewählt wurden (bei 80% der Folds über beide Messzeitpunkte hinweg und Tendenz > $\sqrt{P}$). Es ist jedoch kein Einfluss auf die PRAUC-Werte zu erkennen. Die minimale Anzahl Fälle pro Knoten und somit die Tiefe der einzelnen Entscheidungsbäume unterscheidet sich zwischen den Folds, wobei gemäss PRAUC keine gewählte Tiefe deutliche Vorteile vorweisen kann. Auch der tiefste gewählte Wert weicht vom Default Wert von eins ab und der tiefste übergebene Wert fünf wurde bei nur einem Fold ausgewählt. Beim Betrachten der PRAUC Werte, welche verwendet wurden, um die beste Hyperparameterkombination in jedem Fold auszuwählen, ist zu beobachten, dass diese Werte in etwa doppelt so gross sind, als bei einer zufälligen Schätzung zu erwarten wäre (Ozenne et al., 2015). Dabei fällt auch auf, dass die PRAUC Werte jeweils beim zweiten Messzeitpunkt nur geringfügig höher sind als beim Ersten. Dies kann möglicherweise auf den geringen Zeitabstand von drei Monaten zwischen den Messungen zurückgeführt werden.

Tabelle 6
*Pro Fold ausgewählte Hyperparameter und PRAUC PSYCHE-D*

| Messzeitpunkt | Fold | Regel für die Entscheidung | Anzahl zufällig ausgewählte Variablen für jeden Split | Minimale Anzahl Fälle pro Knoten | PRAUC |
|---|---|---|---|---|---|
| 1 | 1 | Gini | 10 | 30 | .58 |
| 1 | 2 | Hellinger | 6 | 30 | .58 |
| 1 | 3 | Extratrees | 6 | 10 | .58 |
| 1 | 4 | Extratrees | 6 | 20 | .57 |
| 1 | 5 | Extratrees | 12 | 10 | .59 |
| 2 | 1 | Gini | 8 | 40 | .59 |
| 2 | 2 | Hellinger | 6 | 40 | .59 |
| 2 | 3 | Extratrees | 10 | 20 | .60 |
| 2 | 4 | Extratrees | 6 | 5 | .60 |
| 2 | 5 | Gini | 2 | 40 | .60 |

*Anmerkung.* Beim PRAUC gilt zu beachten, dass dieser stark von dem Anteil der positiven Klasse abhängig ist und somit im Verhältnis zu dieser einzuschätzen ist (Ozenne et al., 2015). Beim Datensatz PSYCHE-D beträgt diese 30.4%. Die PRAUC-Werte wurden über die inneren Folds bestimmt anhand von denen das Modell mit den entsprechenden Hyperparametern evaluiert und ausgewählt wurde.

## 4.2. Probabilistische Modelle nach Messzeitpunkt

In diesem Kapitel werden die im Rahmen der Diskriminanzanalyse erstellten Modelle beschrieben. Es handelt sich dabei um die Diskriminanzanalysen, die auf Basis der Ergebnisse der Random Forests pro äusserem Fold erstellt wurden.

**4.2.1. Schweizer Haushalt-Panel.** Die wichtigsten Parameter der Modelle, welche im Rahmen der Diskriminanzanalyse pro Fold beim Datensatz Schweizer Haushalt-Panel erstellt wurden, sind in Tabelle 7 über die Folds hinweg gemittelt eingetragen. Ausserdem ist die Spannweite der Werte über die Folds hinweg angegeben. Es zeigt sich, dass im Prävalenz- sowie im Markermodell das Geschlecht einen Einfluss auf die Schätzung hat, währenddem das Alter oft keinen oder nur geringen Einfluss hat ($b_{Alter}$ zwischen 0.00 und 0.01 in allen Modellen). Die Vergleichbarkeit ist jedoch eingeschränkt, da es sich um unstandardisierte Regressionskoeffizienten handelt (Kuckartz et al., 2013). Dabei ist der Wert für Personen,

welche weibliches Geschlecht angegeben haben, im Prävalenzmodell ($\overline{b}_{Geschlecht:weiblich}$ = 0.42) jeweils höher als für Personen, die männliches Geschlecht angegeben haben. Dasselbe lässt sich beim Markermodell in der Interaktion aus Geschlecht und späterer erhöhter depressiver Symptomatik erkennen, welche einen ersichtlichen Einfluss auf die Schätzung hat ($\overline{b}_{Geschlecht:Spätere\ Symptomatik\ Zeitpunkt\ 1} = -0.27$ und $\overline{b}_{Geschlecht:Spätere\ Symptomatik\ Zeitpunkt\ 2} = -0.27$), währenddem die Interaktion aus Alter und späterer erhöhter depressiver Symptomatik keinen erkennbaren Einfluss zeigt ($\overline{b}_{Alter:Spätere\ Symptomatik\ Zeitpunkt\ 1} = 0.00$ und $\overline{b}_{Alter:Spätere\ Symptomatik\ Zeitpunkt\ 2} = 0.00$). Es ist zu erkennen, dass die Effekte im Prävalenzmodell sowie dem Markermodell über die Folds hinweg in dieselbe Richtung gehen. Auch die Extremwerte der verschiedenen Folds gehen nicht stark auseinander, was dafürspricht, dass die erstellten Modelle über die Folds hinweg ähnlich sind. Werden die Kovarianzen ($\overline{cov}$ = 0.44) zwischen beiden Messzeitpunkten im Verhältnis zu den Residualvarianzen betrachtet, fällt auf, dass die Werte bei den beiden Messzeitpunkten innerhalb von Personen nicht unabhängig voneinander sind, jedoch auch nicht so gross, dass sich die beiden Modelle der unterschiedlichen Messzeitpunkte nicht voneinander abgrenzen. Die daraus abgeleitete mittlere Pearson-Korrelation beträgt $\bar{r}$ = .44 (Kuckartz et al., 2013).

Tabelle 7
*In der Diskriminanzanalyse erstellten Modelle über Folds hinweg Schweizer Haushalt-Panel*

| | Wert | Mittelwert | Minimum | Maximum |
|---|---|---|---|---|
| Prävalenzmodell | | | | |
| | Intercept | -3.29 | -3.47 | -2.97 |
| | Alter | 0.00 | 0.00 | 0.01 |
| | Geschlecht:weiblich | 0.42 | 0.33 | 0.49 |
| | AIC | 1 565 | 1 561 | 1 572 |
| Vorhersage Marker Zeitpunkt 1 | | | | |
| | Intercept | -2.73 | -2.82 | -2.46 |
| | Alter | 0.00 | 0.00 | 0.00 |
| | Geschlecht:weiblich | 0.38 | 0.31 | 0.43 |
| | Spätere Symptomatik:erhöht | 2.50 | 2.05 | 2.84 |
| | Alter:Spätere Symptomatik | 0.00 | 0.00 | 0.00 |
| | Geschlecht:Spätere Symptomatik | -0.27 | -0.35 | -0.17 |
| | Residualvarianz | 0.98 | 0.71 | 1.10 |
| Vorhersage Marker Zeitpunkt 2 | | | | |
| | Intercept | -2.83 | -3.10 | -2.51 |
| | Alter | 0.00 | 0.00 | 0.00 |
| | Geschlecht:weiblich | 0.34 | 0.27 | 0.40 |
| | Spätere Symptomatik:erhöht | 2.66 | 2.10 | 3.14 |
| | Alter:Spätere Symptomatik | 0.00 | 0.00 | 0.00 |
| | Geschlecht:Spätere Symptomatik | -0.27 | -0.35 | -0.11 |
| | Residualvarianz | 1.04 | 0.74 | 1.28 |
| Kovarianz der Residuen zwischen Messzeitpunkten | | 0.44 | 0.37 | 0.54 |

*Anmerkung.* Bei Geschlecht ist jeweils die Referenzkategorie männlich und bei der späteren Symptomatik keine erhöhte depressive Symptomatik. Die Ergebnisse werden so berichtet, dass bei der abhängigen Variable ein höherer Wert einer grösseren Wahrscheinlichkeit an depressiven Symptomen entspricht. Die Werte wurden jeweils für jeden äusseren Fold mit den jeweiligen Trainingsdaten geschätzt und für die Tabelle zusammengefasst. Source: Swiss Household Panel (SHP).

**4.2.2. PSYCHE-D.** Die wichtigsten Parameter der Modelle, welche im Rahmen der Diskriminanzanalyse pro Fold beim Datensatz PSYCHE-D erstellt wurden, sind in Tabelle 8 über die Folds gemittelt eingetragen. Ausserdem ist die Spannweite der Werte über die Folds hinweg angegeben. Es zeigt sich wie beim Datensatz Schweizer Haushalt-Panel im Prävalenz- sowie im Markermodell, dass das Geschlecht einen Einfluss auf die Schätzung hat, währenddem das Alter nur einen geringen Einfluss hat ($b_{Alter}$ zwischen -0.01 und 0.01 in allen Modellen). Die Vergleichbarkeit ist jedoch eingeschränkt, da es sich um unstandardisierte

Regressionskoeffizienten handelt (Kuckartz et al., 2013). Dabei ist der Wert für Personen, welche männliches Geschlecht angegeben haben, im Prävalenzmodell ($\overline{b}_{Geschlecht:männlich}$ = -0.80) jeweils wiederum tiefer als für Personen, die weibliches Geschlecht angegeben haben. Dasselbe lässt sich beim Markermodell in der Interaktion aus Geschlecht und späterer erhöhter depressiver Symptomatik erkennen, welche einen ersichtlichen Einfluss auf die Schätzung hat ($\overline{b}_{Geschlecht:Spätere\ Symptomatik\ Zeitpunkt\ 1} = 0.23$ und $\overline{b}_{Geschlecht:Spätere\ Symptomatik\ Zeitpunkt\ 2} = 0.21$), währenddem die Interaktion aus Alter und späterer erhöhter depressiver Symptomatik keinen erkennbaren Einfluss zeigt ($\overline{b}_{Alter:Spätere\ Symptomatik\ Zeitpunkt\ 1} = 0.00$ und $\overline{b}_{Alter:Spätere\ Symptomatik\ Zeitpunkt\ 2} = 0.00$). Es ist zu erkennen, dass die Effekte im Prävalenzmodell sowie dem Markermodell über die Folds hinweg in dieselbe Richtung gehen. Auch die Extremwerte der verschiedenen Folds gehen nicht stark auseinander, was dafürspricht, dass die erstellten Modelle über die Folds hinweg ähnlich sind. Bei den Residualvarianzen des Markermodells ist zu erkennen, dass die Werte im Vergleich über die Zeitpunkte hinweg ähnlich sind. Die Beobachtung der Kovarianz ($\overline{cov}$ = 0.21) im Verhältnis zur Varianz deutet wiederum daraufhin, dass die Prognosen der Messzeitpunkte innerhalb von Personen zusammenhängen, jedoch nicht so stark, dass die Abgrenzungsfähigkeit in Frage gestellt werden muss. Die daraus abgeleitete mittlere Pearson-Korrelation beträgt $\bar{r}$ = .57 (Kuckartz et al., 2013). Dabei ist diese Trennung beim Datensatz PSYCHE-D weniger klar zu erkennen als beim Schweizer Haushalt-Panel.

Tabelle 8
*In der Diskriminanzanalyse erstellten Modelle über Folds hinweg PSYCHE-D*

| | Wert | Mittelwert | Minimum | Maximum |
|---|---|---|---|---|
| Prävalenzmodell | | | | |
| | Intercept | -0.18 | -0.39 | -0.08 |
| | Alter | -0.01 | -0.01 | -0.01 |
| | Geschlecht:männlich | -0.80 | -0.93 | -0.68 |
| | AIC | 1 687 | 1 679 | 1 697 |
| Vorhersage Marker Zeitpunkt 1 | | | | |
| | Intercept | -1.40 | -1.65 | -1.18 |
| | Alter | 0.01 | 0.00 | 0.01 |
| | Geschlecht:männlich | -0.41 | -0.47 | -0.35 |
| | Spätere Symptomatik:erhöht | 1.90 | 1.44 | 2.33 |
| | Alter:Spätere Symptomatik | 0.00 | -0.01 | 0.00 |
| | Geschlecht:Spätere Symptomatik | 0.23 | 0.14 | 0.29 |
| | Residualvarianz | 0.38 | 0.29 | 0.49 |
| Vorhersage Marker Zeitpunkt 2 | | | | |
| | Intercept | -1.32 | -1.52 | -1.10 |
| | Alter | 0.00 | 0.00 | 0.00 |
| | Geschlecht:männlich | -0.40 | -0.50 | -0.27 |
| | Spätere Symptomatik:erhöht | 1.78 | 1.31 | 2.17 |
| | Alter:Spätere Symptomatik | 0.00 | 0.00 | 0.00 |
| | Geschlecht:Spätere Symptomatik | 0.21 | 0.06 | 0.36 |
| | Residualvarianz | 0.36 | 0.21 | 0.47 |
| Kovarianz der Residuen zwischen Messzeitpunkten | | 0.21 | 0.17 | 0.26 |

*Anmerkung.* Bei Geschlecht ist jeweils die Referenzkategorie weiblich und bei der späteren Symptomatik keine erhöhte depressive Symptomatik. Die Ergebnisse werden so berichtet, dass bei der abhängigen Variable ein höherer Wert einer grösseren Wahrscheinlichkeit an depressiven Symptomen entspricht. Die Werte wurden jeweils für jeden äusseren Fold mit den jeweiligen Trainingsdaten geschätzt und für die Tabelle zusammengefasst.

## 4.3. Sequentielle Selektion Prädiktionsleistung

Folgend werden die Ergebnisse zur Prädiktionsleistung vorgestellt. Dabei werden Prädiktionen verglichen, welche immer nur einen oder immer beide Messzeitpunkte zur Zuteilung verwenden und Prädiktionen, welche eine adaptive sequentielle Selektion der Messzeitpunkte nutzen. Auf Grund der ungleich verteilten Klassen wird zur Bewertung der Prädiktionsleistung die BA verwendet (Hancock et al., 2023; Liu et al., 2023).

**4.3.1. Schweizer Haushalt-Panel.** Wird das Modell mit Schätzungen aus den Random Forests in Abbildung 2 (oben links) betrachtet, welches die Gewichtung der Fehler falsch positiv und falsch negativ gleich und somit den Cutoff für die Zuteilung bei .5 setzt, ist zu erkennen, dass die BA eher gering ist ($BA_{Zwei\ Messungen}$ = .67). Dies ist vor allem auffällig, da das Modell, welches nur die aktuelle depressive Symptomatik zur Prognose nutzt (Ergebnisse aus den logistischen Regressionen mit PC17 und PC18 als Prädiktorvariablen erstellt an der Diagnose Depression), bei derselben Gewichtung der Fehlklassifikationen zu einer leicht höheren BA führt ($BA_{Zwei\ Messungen}$ = .69; Abbildung 2 unten links). Dies kann auf die Probleme zurückzuführen sein, die bei Random Forest Modellen bei stark unausgewogenen Klassen auftreten können, wenn der Cutoff nicht entsprechend angepasst wird. Eine weitere

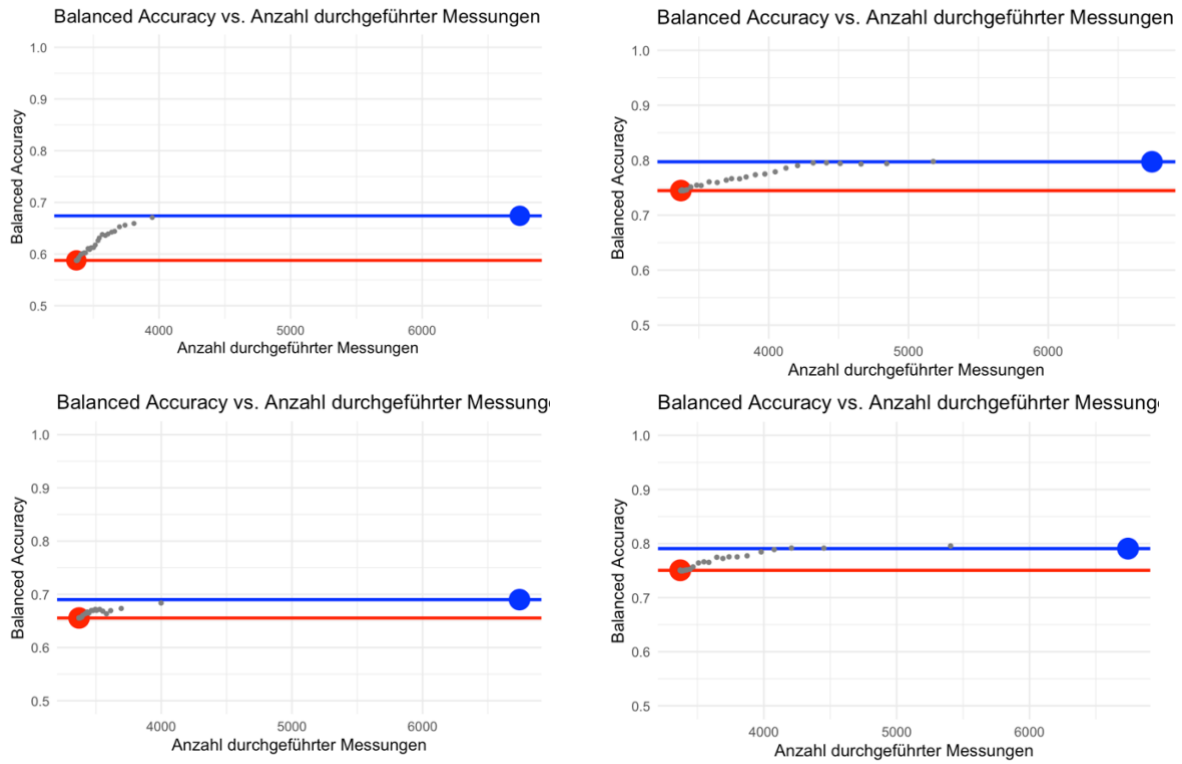


*Abbildung 2.* BA einmalige, zweifache und sequentielle Messung Schweizer Haushalt-Panel. *Anmerkung.* Oben links ist die sequentielle Auswahl mit den Modellen aus den Random Forests abgebildet mit gleicher Gewichtung der Fehler. Rechts davon ist die Auswahl mit den Modellen aus den Random Forests abgebildet mit Gewichtung der Fehler nach inversem Anteil der späteren erhöhten depressiven Symptomatik. Unten links ist die Abbildung, welche nur die Ergebnisse aus der logistischen Regression mit PC17 und PC18 zur Prognose verwendet mit gleicher Gewichtung der Fehler und untern rechts mit Gewichtung der Fehler nach inversem Anteil. Die rote Linie entspricht der BA bei einer Messung, die blaue bei jeweils zwei Messungen und die grauen Punkte zeigen Ergebnisse der sequentiellen Auswahl mit unterschiedlichen Messkosten. Source: Swiss Household Panel (SHP).

mögliche Erklärung wäre Overfitting. In beiden Modellen ist jedoch zu beobachten, dass die Prognose mit beiden Messzeitpunkten zu einer höheren BA führt als die Prognose mit nur einem Messzeitpunkt. Ausserdem führt die sequentielle Auswahl, in Abhängigkeitder variierenden Kosten der Messung und somit der Anzahl ausgewählter zweiter Messungen, in der Tendenz zu einer steigenden Prognosegüte im Vergleich zu nur einer Messung. Dies deutet darauf hin, dass die POSEIDON Algorithmen die korrekten Personen identifizieren, welche bereits nach einer Messung zugeteilt werden können und bei welchen es eine zweite Messung benötigt.

Wird die Gewichtung der Fehler auf den inversen Anteil der später erhöhten depressiven Symptomatik gesetzt (Abbildung 2 oben rechts), ist zu erkennen, dass das Modell, welches auf Basis der Random Forest Prädiktionen erstellt wurde, eine auf zwei Nachkommastellen gerundet ähnlich gute BA aufweist wie das Modell, welches nur mit der aktuellen depressiven Symptomatik zur Prädiktion erstellt wurde ($BA_{\text{Eine Messung Symptome}} = .75$ und $BA_{\text{Eine Messung Random Forest}} = .74$, $BA_{\text{Zwei Messungen Symptome}} = .79$ und $BA_{\text{Zwei Messungen Random Forest}} = .80$). Folglich ist zu erkennen, dass die erstellten Random Forests beim Datensatz Schweizer Haushalt-Panel keine bedeutend bessere Vorhersagegüte erreichen im Vergleich zur Prädiktion nur mit der aktuellen depressiven Symptomatik, die BA durch die komplexen Modelle und Zunahme von Variablen jedoch auch nicht verschlechtert wird.

Bei Betrachtung der Modelle, welche mit Random Forests und mehr Variablen erzeugt wurden (oben Abbildung 2), ist zu erkennen, dass die sequentielle Auswahl bei beiden Vergleichen mit unterschiedlichen Fehlklassifikationsgewichten eine ähnliche BA erreichen, wie die Prognosen, welche immer beide Messungen verwenden und das mit einer deutlichen Reduktion der Anzahl Messungen (beispielsweise Abbildung 2 oben rechts; $BA_{\text{Sequentiell}0.0125} = .80$ und Reduktion der totalen Messungen auf 5 177 und somit um 23.2% im Vergleich zu immer beiden Messungen). Bei definierten Messkosten von 0.2 zeigt sich beispielsweise ein Tradeoff, da die Anzahl Messzeitpunkte weiter reduziert werden kann (um 46.1%) die BA

jedoch auch abnimmt ($BA_{Sequentiell\ 0.2}$ = .76). Es ist ausserdem zu erkennen, dass ab definierten Messkosten von 0.3375 für alle Personen immer nur eine Messung ausgewählt wird bei der sequentiellen Auswahl. Die Ergebnisse sind ausführlich und tabellarisch im Anhang I aufgeführt. Dasselbe ist auch bei den Modellen zu erkennen, welche nur die aktuelle depressive Symptomatik zur Prognose nutzen (beispielsweise unten rechts Abbildung 2 $BA_{Sequentiell0.025}$ = .79 und Reduktion um 37.6% auf 4 208 Messungen).

Die Tendenz, dass mehr zweite Messzeitpunkte zu einer höheren BA führen, ist nur bei der Prognose mit der aktuellen depressiven Symptomatik und gleicher Gewichtung der Fehlklassifikationen weniger klar zu erkennen (Abbildung 2 unten links). So werden bei Messkosten von 0.0375 3 581 Messungen durchgeführt was zu $BA_{Sequentiell0.375}$ = .66 führt, währenddem bei Messkosten von 0.05 3 553 Messungen durchgeführt werden und $BA_{Sequentiell0.05}$ = .67 erzielt wird. Dies kann darauf zurückgeführt werden, dass auf zwei Nachkommastellen gerundet kein Unterschied bezüglich einer starren Messung und zwei starren Messungen bezüglich Spezifität (.97) gefunden werden kann und somit die Verbesserung der BA stark von der Verbesserung der Sensitivität abhängig ist ($Sensitivität_{Eine\ Messung}$ = .34 und $Sensitivität_{Zwei\ Messungen}$ = .41). Dabei kann es auf Grund des geringen Anteils der positiven Klasse und somit weniger Fällen zu Schwankungen in der Sensitivität kommen. Trotzdem kann auch hier gefunden werden, dass in der Tendenz mehr Messungen zu einer höheren BA führen und mit einer deutlichen Reduktion an zweiten Messungen eine ähnliche BA erreicht werden kann wie mit zwei starr ausgewählten Messzeitpunkten, was sich durch eine signifikante positive Spearman Rangkorrelation zwischen Anzahl Messungen und BA zeigt bei der sequentiellen Auswahl mit verschiedenen definierten Messkosten ($r$ = .93, $p$ < .001).

**4.3.2. PSYCHE-D.** Ähnliche Ergebnisse zeigen sich bei der Betrachtung der BA beim Datensatz PSCYHE-D in Abbildung 3. So ist in allen vier Vergleichen zu erkennen, dass die kombinierte Prädiktion mit beiden Messungen zu einer höheren BA führt als bloss mit der

einmaligen Messung. Ausserdem werden bei der sequentiellen Auswahl, wenn mehr zweite Messungen ausgewählt werden wegen geringeren gewählten Kosten der Messung, in der Tendenz bessere BA-Werte erzielt, als wenn weniger zweite Messzeitpunkte ausgewählt werden wegen höheren definierten Messkosten. Dies deutet wiederum daraufhin, dass die Algorithmen die korrekten Personen identifizieren, welche bereits nach einer Messung zugeteilt werden können und bei welchen es eine zweite Messung benötigt. Werden die Modelle verglichen, welche mit den Random Forests erzeugt wurden, mit denen, welche nur über den PHQ-9 Score berechnet wurden, zeigt sich bei beiden Arten der Gewichtung der Fehlklassifikationskosten eine ähnliche maximale BA (bei gleicher Gewichtung $BA_{\text{Zwei Messungen Random Forest}} = .79$, $BA_{\text{Zwei Messungen Symptomatik}} = .78$ und bei der Gewichtung nach inversem Anteil $BA_{\text{Zwei Messungen Random Forest}} = .79$, $BA_{\text{Zwei Messungen Symptomatik}} = .81$). Dies deutet wiederum darauf hin, dass die erstellten Random Forest Modelle und die zusätzlichen Variablen die Prognosegüte nicht verbessern können im Vergleich zur Prognose nur mit dem PHQ-9 Score, diese jedoch auch nicht verschlechtern.

Bei den Modellen, die auf Bais der Random Forests erstellt wurden, ist zu erkennen, dass sich die BA bei nur einer durchgeführten Messung deutlich verbessert, wenn statt der gleichen Kosten für Fehlklassifikationen die gewichtete nach inversem Anteil verwendet wird, währenddem sich die BA bei zwei Messungen kaum verändert. Dies führt dazu, dass bei der Gewichtung der Kosten der Fehlklassifikation nach inversem Anteil der Unterschied in der BA zwischen einer und zwei Messungen nur sehr gering ist und somit nur wenig Spielraum für die sequentielle Auswahl bleibt ($BA_{\text{Eine Messung}} = .78$, $BA_{\text{Zwei Messungen}} = .79$). Dies zeigt sich, da die Anpassung des Cutoffs einen stärkeren Effekt auf die Prädiktionsleistung hat, wenn die Prädiktion eine suboptimale Trennschärfe aufweist, als wenn die Klassen bereits gut trennbar sind (Fawcett, 2006).

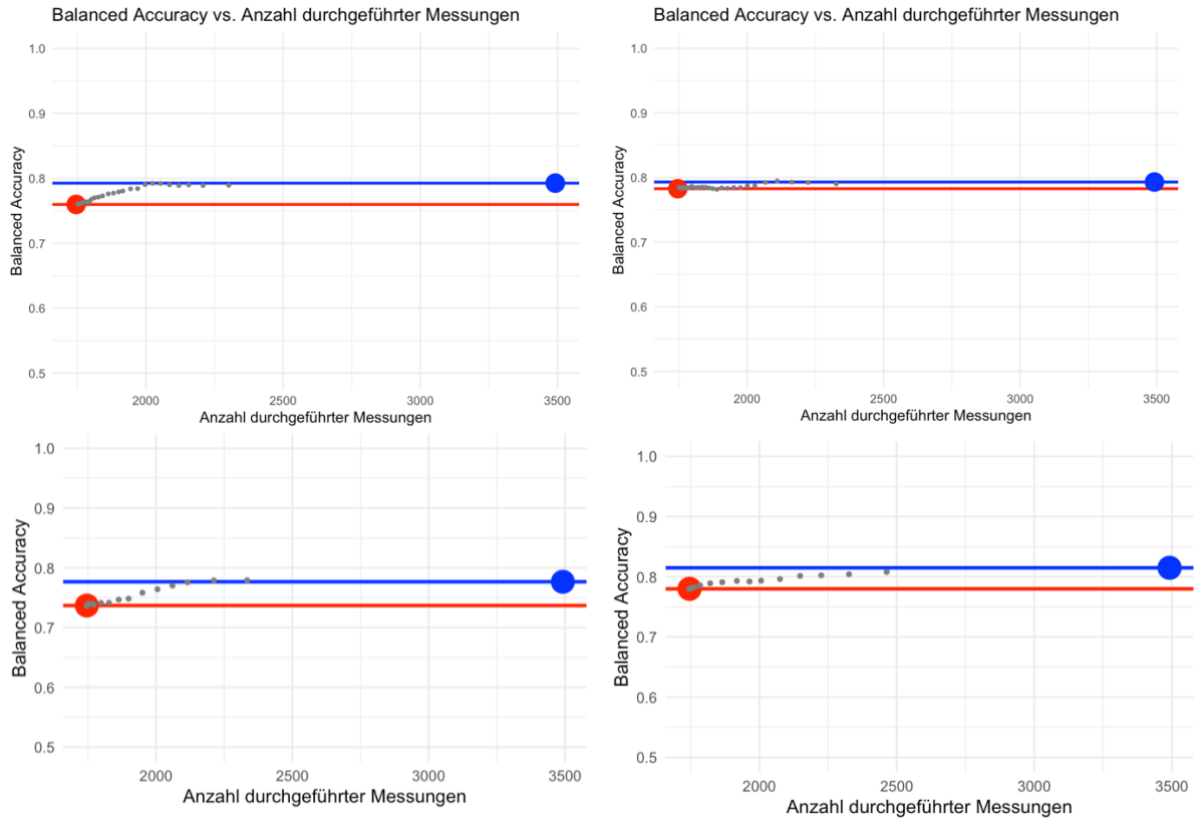


*Abbildung 3.* BA einmalige, zweifache und sequentielle Messung PSYCHE-D. *Anmerkung.* Oben links ist die sequentielle Auswahl mit den Modellen aus den Random Forests abgebildet mit gleicher Gewichtung der Fehler. Rechts davon ist die Auswahl mit den Modellen aus den Random Forests abgebildet mit Gewichtung der Fehler nach Anteil der späteren erhöhten depressiven Symptomatik. Unten links ist die Abbildung, welche nur die PHQ-9 Werte zur Prognose nutzte mit gleicher Gewichtung der Fehler und untern rechts mit Gewichtung der Fehler nach inversem Anteil. Die rote Linie entspricht der BA bei einer Messung, die blaue bei jeweils zwei Messungen und die grauen Punkte zeigen Ergebnisse der sequentiellen Auswahl mit unterschiedlichen Messkosten.

Wird die Accuracy zur Prognosegenauigkeit in Abbildung 4 betrachtet, wird weiter verdeutlicht, dass die Unterschiede in der Prognosegüte von einer starren und zwei starren Messung nur gering sind bei der Gewichtung der Fehlklassifikationskosten nach inversem Anteil bei der Prognose mit Random Forest. Dabei wird jedoch verdeutlich, dass durch eine Zunahme der ausgewählten zweiten Messzeitpunkte wegen geringer definierten Messkosten die Prognosegenauigkeit steigt. Es kann auch wiederum gezeigt werden, dass mit weniger Messungen eine vergleichbare BA gezeigt werden kann, indem sequentiell Messungen ausgewählt werden, im Vergleich zur starren Auswahl von zwei Messungen.

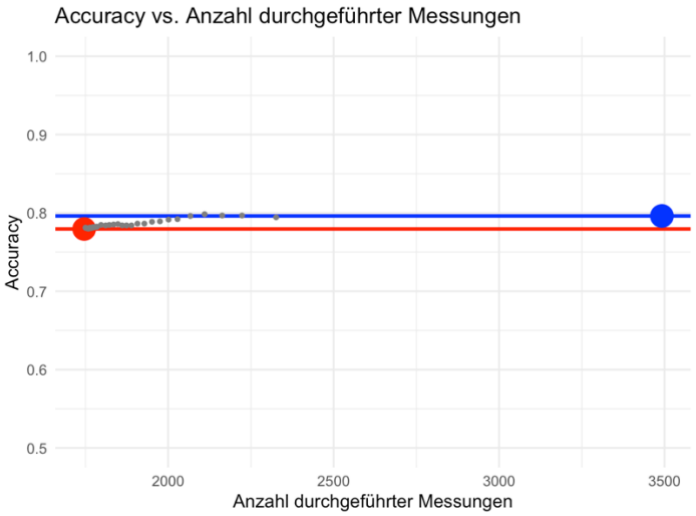


*Abbildung 4.* Accuracy einmalige, zweifache und sequentielle Messung PSYCHE-D Gewichtung nach inversem Anteil der späteren erhöhten depressiven Symptomatik. *Anmerkung.* Die rote Linie entspricht der Accuracy bei einer Messung, die blaue bei jeweils zwei Messungen und die grauen Punkte zeigen Ergebnisse der sequentiellen Auswahl mit unterschiedlichen Messkosten.

Folgend wird bei den Random Forest Modellen die Selektion mit der Gewichtung der Fehlklassifikationskosten nach inversem Anteil betrachtet, da dieses mit zwei Messungen eine leicht höhere BA vorweist als die Selektion mit gleicher Gewichtung (siehe Abbildung 3 oben rechts). So führt bei definierten Messkosten von 0.0125 die sequentielle Auswahl im Vergleich zur starren Auswahl von zwei Messzeitpunkten zu einer Reduktion um 1 166 Messungen (33.4%), wobei die BA auf zwei Nachkommastellen gerundet gleichbleibt ($BA_{Sequentiell0.0125}$ = .79). Bei Kosten von 0.0875 für die Messung kann eine Reduktion um 1 391 (42.7%) Messungen erreicht werden, wobei die BA weiterhin gleich bleibt auf zwei Nachkommastellen gerundet. Bei Messkosten von 0.1 zeigt sich ein Trade off, da die Anzahl an Messungen weiterhin gesenkt werden kann (um 43.4%), die BA jedoch auch abnimmt ($BA_{Sequentiell0.1}$ = .78). Die Ergebnisse sind ausführlich und tabellarisch im Anhang J aufgeführt.

Ähnliches zeigt sich bei der Betrachtung der BA bei den Modellen, welche nur mit der aktuellen depressiven Symptomatik erzeugt wurden. Auch dort führt die Gewichtung der Fehlklassifikationskosten nach inversem Anteil und entsprechender Anpassung des Cutoffs zu einer leicht höheren BA als bei gleicher Gewichtung der Fehlklassifikationskosten. So führen

definierte Messkosten von 0.0125 zu $BA_{Sequentiell0.0125}$ = .81 und zu einer Reduktion auf 2 463 Messungen (29.5% Reduktion im Vergleich zu zwei starren Messzeitpunkten). Dabei zeigt sich beispielsweise ein Tradeoff bei definierten Messkosten von 0.0625, da weniger zweite Messungen durchgeführt werden, die BA im Vergleich zu zwei starren Messzeitpunkten jedoch auch abnimmt ($BA_{Sequentiell0.0625}$ = .80 bei Reduktion der Messungen auf 2 075 und somit um 40.6% im Vergleich zu zwei starren Messzeitpunkten).

## 4.4. Sequentielle Selektion Kosten

Folgend werden die Ergebnisse zu den durchschnittlichen Kosten nach Selektionsmethode vorgestellt. Dabei werden Prädiktionen verglichen, welche immer nur einen oder zwei Messzeitpunkte zur Zuteilung verwenden und Prädiktionen, welche eine adaptiv sequentielle Selektion der Messzeitpunkte nutzen. Die durchschnittlichen Gesamtkosten pro Person ergeben sich aus den Messkosten und den Fehlklassifikationskosten. Es werden nur die Modelle betrachtet, welche Random Forests zur Prädiktion verwendet haben, da diese eine ähnliche Prädiktionsleistung erreicht haben wie die Modelle, welche nur mit früherer Messung der depressiven Symptomatik erstellt wurden und bislang im Rahmen dieser Arbeit fokussiert wurden (siehe Abbildungen 2 und 3). Übersichten der Werte sind in Anhang K für den Datensatz Schweizer Haushalt-Panel und Anhang L für den Datensatz PSYCHE-D zu finden. Die Ergebnisse aus der Anwendung der Kostenverteilung mit der Prädiktion nur mit der aktuellen Symptomatik ist im Anhang M für beide Datensätze abgelegt. Dabei zeigen sich im Vergleich der Modelle, welche nur mit der depressiven Symptomatik zur Prädiktion erstellt wurden und der Modelle, welche mit einer grösseren Anzahl Variablen und Random Forests erstellt wurden, Ähnlichkeiten in der relativen Kostenverteilung von einem starren, zwei starren und sequentiell ausgewählten Messzeitpunkten.

**4.4.1. Schweizer Haushalt-Panel.** In Abbildung 5 sind die durchschnittlichen Kosten pro Person graphisch abgebildet. Es ist visuell zu erkennen, dass die sequentielle Auswahl,

besonders bei steigenden Kosten für die Messung, deutlich weniger durchschnittliche Kosten verursacht als die starre Auswahl von zwei Messzeitpunkten. Beim Vergleich der sequentiellen Auswahl mit der starren Auswahl von einem Messzeitpunkt ist diese Tendenz weniger klar zu erkennen. Werden die durchschnittlichen Gesamtkosten bei der gleichen Gewichtung der Fehlklassifikationskosten betrachtet (Abbildung 5 links), ist zu erkennen, dass bei geringen definierten Kosten der Messung die sequentielle Auswahl weniger Kosten im Durchschnitt verursacht. So hat die sequentielle Auswahl bei definierten Messkosten von 0.0125 durchschnittliche Kosten von $M_{Kosten\ Sequentiell0.0125} = 0.091$, währenddem die Prädiktion mit starr einer Messung Kosten von $M_{Kosten\ Eine\ Messung0.0125} = 0.096$ hat (Reduktion um 5.2%). Auch bei gewählten Messkosten bis 0.05 ist diese Tendenz weiterhin zu erkennen (Reduktion um 1.5%). Dabei nähern sich die durchschnittlichen Kosten der sequentiellen Auswahl bei steigenden definierten Messkosten an die der einmaligen Messung an und übersteigen diese sogar leicht bei zunehmenden definierten Messkosten, maximal um 1.6%. Ab Messkosten von 0.3375 werden die Kosten von einer starren und sequentiell ausgewählten Messzeitpunkten identisch, da auch bei der sequentiellen Auswahl immer nur die erste Messung ausgewählt wird.

Werden die Kosten für die Fehlklassifikation nach inversem Anteil gewichtet (Abbildung 5 rechts), ist zu erkennen, dass bei geringen Messkosten von 0.0125 und 0.025 die durchschnittlichen Kosten bei der sequentiellen Auswahl am tiefsten ist, gefolgt von den Prädiktionen mit zwei Messungen und zuletzt den Prädiktionen mit nur einer Messung (bei Messkosten von 0.0125 aufsteigend $M_{Kosten\ Sequentiell0.0125} = 0.222$, $M_{Kosten\ Zwei\ Messungen0.0125} = 0.229$, $M_{Kosten\ Eine\ Messung0.0125} = 0.269$ und bei Messkosten von 0.025 aufsteigend $M_{Kosten\ Sequentiell0.025} = 0.243$, $M_{Kosten\ Zwei\ Messungen0.025} = 0.254$, $M_{Kosten\ Eine\ Messung0.025} = 0.281$). Dies entspricht einer Kostensenkung im Vergleich zu einer starren ausgewählten Messung von 17.5% und 13.5%. Bei weiter steigenden

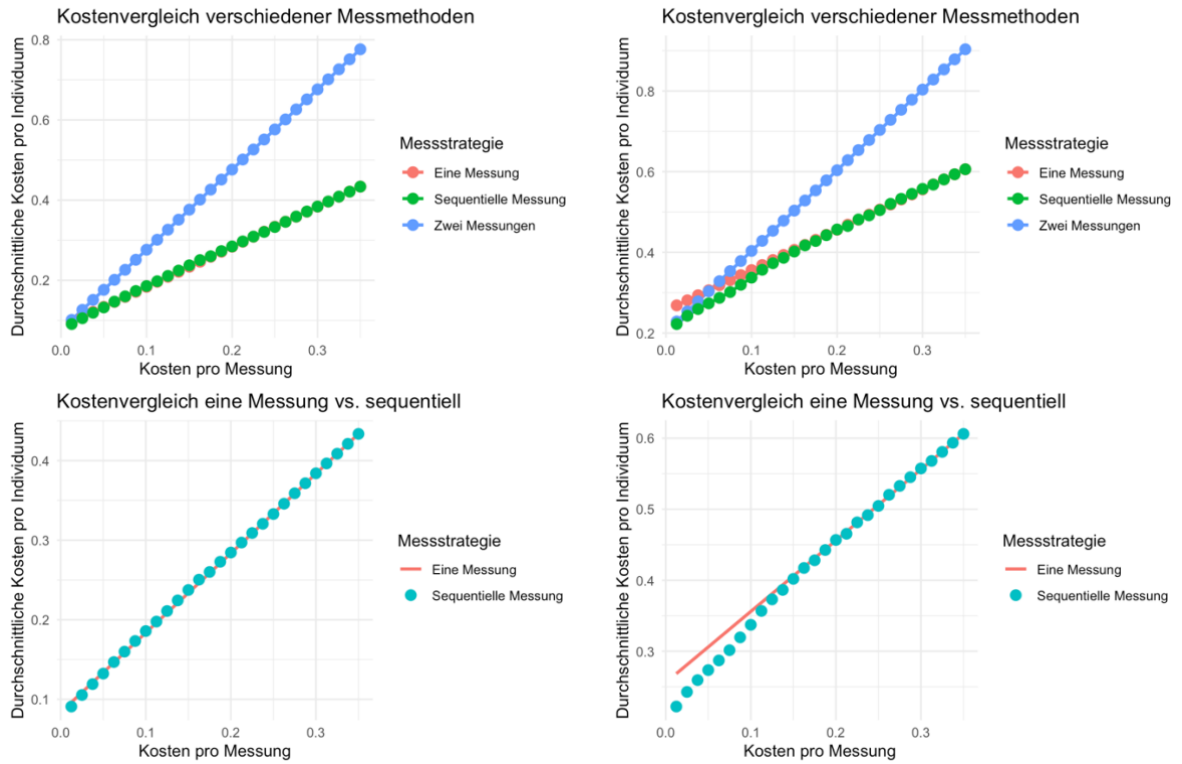


*Abbildung 5.* Totale Kosten einmalige, zweifache und sequentielle Messung Schweizer-Haushalt-Panel. *Anmerkung.* Links sind die gleichen Kosten für die Fehlklassifikation und rechts gewichtete Kosten nach inversem Anteil. Bei den oberen Messungen sind die Kosten für Kosten einmalige, zweifache und sequentielle Messung angegeben, währenddem bei den unteren Graphiken nur die Werte der sequentiellen Messungen, mit denen der einmaligen Messung dargestellt werden, um die Vergleichbarkeit zu verbessern. Source: Swiss Household Panel (SHP).

Messkosten ist zu erkennen, dass die sequentielle Auswahl und die Prädiktion mit einem starren Messzeitpunkt deutlich weniger durchschnittliche Kosten verursachen als die Prädiktion mit zwei starren Messungen. Dabei ist auch zu erkennen, dass bei steigenden Messkosten die einmaligen Messungen etwas weniger durchschnittliche Gesamtkosten verursachen als die sequentiell ausgewählten Messzeitpunkte (maximal 0.4%).

**4.4.2. PSYCHE-D.** Werden die Messkosten beim Datensatz PSYCHE-D in Abbildung 6 (oben links) betrachtet, ist bei der gleichen Gewichtung der Fehlklassifikationskosten zu erkennen, dass die sequentiellen Prognosen bei niedrigen Kosten am besten abschneiden, sich aber bei steigenden Messkosten an die Kosten der starren einzelnen Messung annähern. Prognosen mit zwei starren Messzeitpunkten führen bei steigenden definierten Messkosten zu

deutlich höheren durchschnittlichen Gesamtkosten als die anderen beiden Prognosen mit einmaliger und sequentieller Auswahl. Dabei verursachen die Prognosen mit zwei starren Messzeitpunkten bei gering definierten Messkosten weniger durchschnittliche Kosten als die Prognosen mit einem starren Messzeitpunkt, jedoch höhere als die Prognosen mit sequentieller Auswahl. Bei tiefen definierten Messkosten führen die sequentiellen Prognosen wiederum zu den geringsten durchschnittlichen Gesamtkosten (sequentielle Auswahl hat bei definierten Messkosten von 0.0125 durchschnittliche Gesamtkosten von $M_{Kosten\ Sequentiell0.0125} = 0.203$, zwei starre Messungen von $M_{Kosten\ Zwei\ Messungen0.0125} = 0.209$ und eine starre Messung von $M_{Kosten\ Eine\ Messung0.0125} = 0.226$ und bei Messkosten von 0.025 durchschnittliche Gesamtkosten in derselben Reihenfolge von $M_{Kosten\ Sequentiell0.025} = 0.219$, $M_{Kosten\ Zwei\ Messungen0.025} = 0.234$ und $M_{Kosten\ Eine\ Messung0.025} = 0.238$). Dies entspricht einer Kostensenkung bei der sequentiellen Auswahl im Vergleich zur starren einmaligen Auswahl von 10.2% und 8.0%. Bei steigenden definierten Messkosten nähern sich die durchschnittlichen Gesamtkosten der sequentiellen Auswahl an die der einmaligen Messung an und überschreiten diese teilweise leicht (maximal 0.4%). Ab definierten Messkosten von 0.425 wird bei den sequentiellen Prognosen immer nur ein Messzeitpunkt pro Person ausgewählt.

Bei den Anwendungen, welche die Fehlklassifikation nach inversem Anteil gewichten beim Datensatz PSYCHE-D, ist zu erkennen (Abbildung 6 rechts), dass bei Messkosten von 0.0125 die durchschnittlichen Kosten bei den sequentiellen Prädiktionen ($M_{Kosten\ Sequentiell0.0125} = 0.227$) tiefer sind als bei den Prädiktionen mit starrer Auswahl an Messzeitpunkten (Kosten bei zwei Messzeitpunkten $M_{Kosten\ Zwei\ Messungen0.0125}$ = 0.233 und bei einem Messzeitpunkt $M_{Kosten\ Eine\ Messung0.0125} = 0.231$). Diese Unterschiede sind jedoch gering (Kostensenkung von 1.7% im Vergleich zu einem starren Messzeitpunkt). Ausserdem ist visuell zu erkennen, dass die durchschnittlichen Kosten bei den sequentiellen Messungen

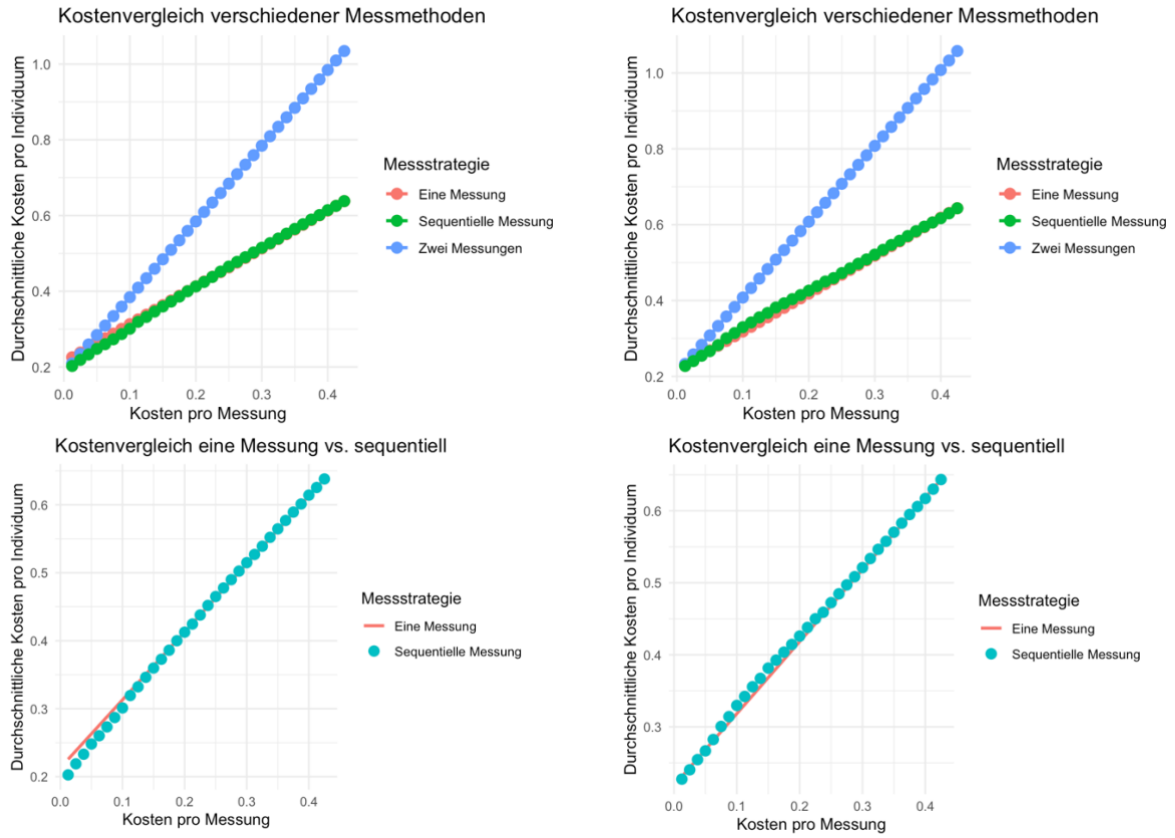


*Abbildung 6.* Totale Kosten einmalige, zweifache und sequentielle Messung PSYCHE-D. *Anmerkung.* Links sind die gleichen Kosten für die Fehlklassifikation und rechts gewichtete Kosten nach inversem Anteil. Bei den oberen Messungen sind die Kosten für Kosten einmalige, zweifache und sequentielle Messungen angegeben, währenddem bei den unteren Graphiken nur die Werte der sequentiellen Messungen, mit denen der einmaligen Messung dargestellt werden, um die Vergleichbarkeit zu verbessern.

diejenigen der einzelnen Messungen bei gewissen definierten Messkosten übersteigen. Dies kann unter anderem darauf zurückgeführt werden, dass bei der Gewichtung der Fehlklassifikationen nach inversem Anteil bei nur einer Messung eine ähnliche BA erreicht wird wie bei zwei Messungen (siehe Abbildungen 3 und 4). Dabei übersteigen die Kosten der sequentiell ausgewählten Messzeitpunkte die der starren einmaligen Messung um maximal 3.8%. Dies kann wiederum auf den kurzen Zeitintervall von drei Monaten zwischen den beiden Messungen beim Datensatz PSYCHE-D zurückgeführt werden. Der daraus resultierende geringe Mehrwert der zweiten Messung erschwert der sequentiellen Auswahl tiefere durchschnittliche Kosten zu verursachen als die einmalige Messung. Dabei ist zu beobachten, dass die zweite Messung zu keiner Verbesserung der Sensitivität führt auf zwei Nachkommastellen gerundet ($Sensitivität_{Eine\ Messung} = .79$, $Sensitivität_{Zwei\ Messungen} = .79$). Dies

wird dadurch verschärft, dass die falsch negative Klassifikation bei der Gewichtung nach inversem Anteil besonders stark gewichtet wird. Bei Messkosten ab 0.425 wird bei der sequentiellen Auswahl schliesslich immer nur eine Messung pro Person ausgewählt.

Folglich führt die sequentielle Auswahl im Vergleich zu zwei starren Messzeitpunkten zu einer Kostensenkung bei allen definierten Messkosten, jedoch besonders deutlich bei steigenden definierten Messkosten. Dahingegen zeigten sich Kostenersparnisse gegenüber einem starren Messzeitpunkt besonders bei tiefen definierten Messkosten. Bei steigenden Messkosten nähern sich die durchschnittlichen Messkosten bei der sequentiellen Auswahl an die der starren einmaligen Messung an und übersteigen diese teilweise auch leicht, insbesondere beim Datensatz PSYCHE-D. Dies ist unter anderem darauf zurückzuführen, dass die Unterschiede in der BA bei zwei Messzeitpunkten im Vergleich zu nur einem gering sind, was auf den kurzen Zeitabstand zwischen den Messungen zurückgeführt werden kann. Im Anhang O ist beispielsweise zu erkennen, dass bezüglich Kosten beim Datensatz PSYCHE-D klarere Vorteile der sequentiellen Auswahl zu erkennen sind im Vergleich zur starren einmaligen Messung bei der Gewichtung der Fehlklassifikationskosten nach Anteil, wenn die Messzeitpunkte neun statt drei Monate auseinanderliegen. Dies kann auf eine grössere Steigerung der BA durch die zweite Messung erklärt werden. Im Anhang N sind Ergebnisse mit zwei Jahren und damit grösserem Zeitabstand beim Datensatz Schweizer Haushalt-Panel abgelegt.

# 5. Diskussion

Im folgenden Teil der Arbeit werden die Ergebnisse in einem ersten Schritt zusammengefasst, theoretisch eingebettet und kritisch diskutiert sowie die Forschungsfragen beantwortet, wobei auch praktische Implikationen diskutiert werden. Daraufhin werden die Limitationen dieser Arbeit genannt und abschliessend wird ein Fazit gezogen und dabei ein Ausblick für weitere Forschung gegeben.

## 5.1. Zusammenfassung

In dieser Arbeit konnten Vorteile der sequentiellen Auswahl von Messzeitpunkten im Vergleich zur starren Auswahl gefunden werden in Abhängigkeit vom definierten Kontext. Jedoch muss erwähnt werden, dass die zusätzlichen Variablen und Random Forest Modelle keine bessere Prognosegüte, gemessen an der BA, gegenüber nur der aktuellen depressiven Symptomatik zur Prädiktion erreichen konnten. Dies ist aus Sicht des multimodalen Prädiktionsansatzes überraschend, gerade da Hyperparametertuning mit nested Cross Validation, automatisierte Variablenselektion und Massnahmen für unbalancierte Klassen bei Random Forests berücksichtigt wurden (Bartz 2023; Degenhardt et al., 2019; Hancock et al., 2023; Hankin & Abela, 2005; Kaplan et al., 1987; Liu et al., 2019; Liu et al., 2023; McEwen, 1998b; Wainer & Cawley, 2021). Dabei wurde jedoch im Theorieteil schon darauf verwiesen, dass zusätzliche Variablen bei kurzem Zeitabstand oft nur wenig inkrementelle Varianz erklären können (Andersson et al., 2023; Christodoulou et al., 2019; Keenan et al., 2009). Im Folgenden werden in einem ersten Schritt mögliche Gründe für die Befunde diskutiert, wobei auch die zur Prädiktion ausgewählten Variablen theoretisch eingebettet werden. Darauffolgend wird auf die Beantwortung der Fragestellung eingegangen.

**5.1.1. Prädiktiver Nutzen betrachteter Variablen gegenüber symptombasierter Vorhersage.** In der vorliegenden Arbeit konnte gefunden werden, dass sich die im Theorieteil betrachteten Variablen zur Prognose der depressiven Symptomatik eignen, da sie durch Boruta ausgewählt wurden und somit eine systematische Trennung ermöglichen (Degenhardt et al., 2019; Kursa & Rudnicki, 2010). Dass sich diese Variablen zur Prädiktion der künftigen depressiven Symptomatik eignen, ist weiterführend im Anhang P zu erkennen. Dabei wurden für die Prognose dieselben Personen und Variablen bis auf die aktuelle depressive Symptomatik verwendet wie in der Hauptuntersuchung der Arbeit. Beim Datensatz Schweizer Haushalt-Panel konnte eine ähnliche Prognosegüte anhand der BA erreicht sowie die Vorteile der zeitlich sequentiellen Auswahl bestätigt werden. Beim Datensatz PSYCHE-D ist zu

beobachten, dass die Prognosegenauigkeit, gemessen an der BA, deutlich abnimmt ohne die aktuelle depressive Symptomatik zur Prognose und die Vorteile des zweiten Messzeitpunkts gering sind. Trotzdem bleibt die Prognosegüte, gemessen an der BA, höher als bei einer zufälligen Prädiktion erwartet. Diese Modelle könnten genutzt werden, wenn keine Messungen zur Symptomatik vorliegen oder erfasst werden können (zum Beispiel in einem beruflich organisationalen Kontext), was besonders bei den passiv mit Wearables gemessenen Daten eine kosteneffiziente Alternative wäre für ein frühzeitiges Monitoring (Ruiz & Kovacs, 2021). Diese Vorhersage ist vermutlich über die aktuelle Ausprägung der Symptomatik vermittelt, was zur Konfundierung führt, dass die Richtung des Zusammenhangs nicht eindeutig bestimmt werden kann und die Gefahr von Zirkularität besteht (Chancellor, Feuston & Chang, 2023). Folglich empfiehlt sich aus Sicht des Ockhams Rasiermessers, zur Prognose nur die zeitlich sequentielle Messung der aktuellen depressiven Symptomatik zu verwenden, bis Variablen identifiziert und Modelle entwickelt wurden, welche zu einer genaueren Prognose führen als bloss die aktuelle depressive Symptomatik (Domingos, 1999). Ein Vorteil der Erfassung mit zusätzlichen Variablen, beispielsweise zum Lebensstil und Arbeitsleben, bietet hingegen die Möglichkeit, in künftiger Forschung und Anwendung individuellen Handlungsbedarf abzuleiten und individuelle Handlungsempfehlungen zu geben (wie beispielsweise durch das ActIv-Projekt bezüglich Schlaf, Ernährung und Bewegung; Deutsche Sporthochschule Köln, 2025).

Beim Datensatz PSYCHE-D konnte gefunden werden, dass Variablen, welche aus den mit Wearables gemessenen Schritt- und Schlafdaten abgeleitet wurden, zur Prognose der zukünftigen depressiven Symptomatik im kritischen Bereich durch Boruta ausgewählt wurden (Degenhardt et al., 2019; Kursa & Rudnicki, 2010). Dies kann bezüglich Schrittdaten unter anderem motivational mit der Selbstbestimmungstheorie erklärt werden (Deci & Ryan, 1985; Ryan et al., 2009). Sowie bezüglich Schritt- und Schlafdaten auch über physikalische Wirkmechanismen, wobei besonders die HHNA in Verbindung mit dem Konzept der

Allostase als zentraler Wirkmechanismus gilt (Archer et al., 2014; Asarnow, 2020; McEwen, 1998a; Rao et al., 2009).

Beim Datensatz Schweizer Haushalt-Panel wurden hingegen Variablen aus weniger Kategorien durch Boruta ausgewählt als gemäss multimodalen Prädiktionsansatz, vorgestellter Theorie und dem Diathese-Stress-Modell zu erwarten gewesen wäre (Hoyer & Knappe, 2020; Kaplan et al., 1987; Liu et al., 2019). So wurden mit der automatisierten Variablenselektion keine der Variablen aus den Kategorien Netzwerk, Lebensereignisse und Haushalt ausgewählt und auch beim Arbeitsleben nur die Variable zur Anzahl Arbeitsstunden in der Woche, was über die aktuelle Symptomatik vermittelt sein könnte. Es fällt auf, dass die ausgewählten Variablen beim Schweizer Haushalt-Panel besonders aus der Kategorie Gesundheit stammen. Bei den Gesundheitsdaten ist zu beobachten, dass die ausgewählten Variablen stark auf die emotionale Wahrnehmung und das emotionale Empfinden eingehen, was ein zentraler Aspekt depressiver Störungen darstellt und auch empirisch als Prädiktor bestätigt ist (Morris et al., 2009; WHO, 2024). Wie gemäss Diathese-Stress-Modell zu erwarten gewesen ist, war auch die Wahrnehmung von Stress ein ausgewählter Prädiktor für die Prognose von psychischen Störungen (Hoyer & Knappe, 2020). Überraschend mit Blick auf die Theorien Job-Demand-Control und Job-Demands-Resources ist, dass nur wenige Aspekte des Arbeitslebens, welche als Ressourcen oder Stressoren wahrgenommen werden können, zur Prognose verwendet wurden (Demerouti et al., 2001; Karasek, 1979). Dabei ist jedoch im Schnittstellenbereich zwischen Arbeit und Freizeit zu erkennen, dass die entsprechenden Variablen einen Einfluss zu haben scheinen, da Variablen, welche zeitbasierte und belastungsbasierte Konflikte beschreiben, zur Prädiktion ausgewählt wurden (Tsaur et al., 2012; Tsaur & Yen, 2018). Ausserdem ist zu beobachten, dass bei der Vorhersage der Entwicklung einer depressiven Störung im Anhang A mehr Variablen zum Arbeitsleben als relevant klassifiziert wurden. Überraschend ist dagegen, dass keine Variablen aus den Kategorien Netzwerk und Haushalt ausgewählt wurden. Eine mögliche Erklärung ist dabei,

dass Fragen zum Haushalt nicht von jeder Person selbst beantwortet wurden, sondern nur von einer Person im Haushalt, wobei unterschiedliche Wahrnehmungen der Situation die Ergebnisse verzerren können (Bertrand & Mullainathan, 2001). Trotzdem ist überraschend, dass keine Variablen ausgewählt wurden, die Hinweise auf den sozialökonomischen Status liefern (Letourneau et al., 2013; WHO, 2017). Bei den Fragen zum Netzwerk könnte eine mögliche Erklärung sein, dass keine Variablen zum ganzheitlich aggregierten Netzwerk erhoben wurden, sondern vermehrt auf einzelne Beziehungen eingegangen wurde (Groot, van den Brink & van Praag, 2007). Ausserdem wurden viele Variablen zum Netzwerk nur alle drei Jahre erhoben und konnten somit nicht verwendet werden (siehe Anhang C). Auch zum Lebensstil wurde im Rahmen der Gesundheit keine Variable zur Prädiktion ausgewählt, was überraschend ist, da Variablen zum Lebensstil und Gesundheitsverhalten einen metaanalytisch nachgewiesenen Effekt auf die Entwicklung von psychischen Störungen haben (Cross-Denny & Robinson, 2017; Zhang et al., 2018). Auch beim Datensatz PSYCHE-D wurden nur wenige Variablen zum Lebensstil zur Prädiktion ausgewählt (bloss zur Veränderung der Einnahme von Medikamenten und anderen Therapien, welche eine geringe Variablenwichtigkeit aufweisen).

Dabei führte die Hinzunahme von Variablen neben den aktuellen depressiven Symptomen zu keiner besseren BA. Dies wäre gemäss dem multimodalen Prädiktionsansatz zu erwarten gewesen, da gemäss diesem multimodale Prognosen besser abschneiden als bloss symptombasierte klinische Fragebögen im Rahmen der mentalen Gesundheit (Kaplan et al., 1987; Liu et al., 2019). Dies könnte unter anderem darauf zurückzuführen sein, dass die abhängige Variable über die angegebene Symptomatik und nicht eine Diagnose gebildet wurde oder auf den geringen Zeitabstand zwischen den Messungen. Letzteres wird dadurch verschärft, dass Stressoren und Ressourcen gemäss Diathese-Stress-Modell und dem Konzept der Allostase langfristig wirken (Hankin & Ablela, 2005; McEwen, 1998b). Dies könnte ein Hinweis dafür sein, dass die aktuelle Ausprägung der erhobenen Variablen stark die aktuelle

Symptomatik vorhersagt, was zu wenig zusätzlichem Informationsgewinn gegenüber der Prädiktion mit der aktuellen Symptomatik führt (Hankin & Ablela, 2005; McEwen, 1998b). Dabei gilt zu beachten, dass die Angaben pro Person und Zeitraum nur einmal beantwortet und nicht etwa über den Zeitraum aggregiert wurden. Dies kann über stimmungskongruente Erinnerung zu einem Response Bias führen und somit zu einem erhöhten Zusammenhang der aktuellen Symptomatik und anderen erfassten Variablen (Anhang M; Fiedler, Nickel, Muehlfriedel & Unkelbach, 2001; Matt, Vazquez & Campbell, 1992). Weiterer Bias in den Antworten könnte durch die wiederholte Befragung derselben Personen vermittelt über Antworttendenzen und Erinnerungseffekte auftreten (Tourangeau, Rips & Rasinski, 2000; Watson & Wooden, 2009). Ausserdem wurden zur Prädiktion keine psychologisch validierten Messungen verwendet. Eine weitere mögliche Erklärung wäre Overfitting der Random Forests, obwohl nested Cross Validation angewendet wurde und die Random Forests zu einer guten Prädiktionsleistung führen (Barreñada et al., 2024; Wainer & Cawley, 2021). Dies kann nicht ausgeschlossen werden, da keine Vergleiche der Prädiktionsleistung zwischen Test- und Trainingsdaten gemacht wurden.

Diese Aspekte wären in künftigen Arbeiten zu prüfen, welche Machine Learning Modelle zur Prädiktion der depressiven Symptomatik nutzen. So konnte in früheren Arbeiten gefunden werden, dass die aktuelle depressive Symptomatik ein starker Prädiktor ist, jedoch inkrementelle Varianz durch zusätzliche validierte psychologische Messungen erklärt werden kann, diese bei kurzem Zeitabstand jedoch häufig nur gering ist (Andersson et al., 2023; Andrews, Qian & Valentine,, 2002; Christodoulou et al., 2019; Hatton et al., 2019; Keenan, Feng, Hipwell & Klostermann, 2009; Rude, Wenzlaff, Gibbs, Vane & Whitney, 2002; 2002; Zetsche & Joormann, 2011). Dabei wird in aktueller Forschung diskutiert, inwiefern Machine Learning Modelle im Vergleich zur Regression bei klinischen Modellen und Zuteilungen einen Mehrwert bieten (Christodoulou et al., 2019). Dies betont die Relevanz künftiger Forschung mit psychologisch validierten Messungen und Machine Learning sowie

Regressionsmodellen inkrementelle Varianz gegenüber der aktuellen depressiven Symptomatik zu erklären für die Entwicklung eines Monitoringinstruments für die zeitlich sequentiell adaptive Anwendung. Die in dieser Arbeit theoretisch hergeleiteten Kategorien können dabei eine Basis bieten.

Ein weiterer wichtiger Aspekt bei der Beurteilung der Modelle bei Machine Learning Verfahren ist die Betrachtung der Hyperparameter (Rimal et al., 2024; Scornet, 2017; Valavi et al., 2021). Auch bei dieser Untersuchung wurde ein Hinweis dafür gegeben, dass sich die Default-Werte oftmals nicht für das Ziel der Untersuchung eignen, da die mit Hyperparametertuning ausgewählten Hyperparameter von den Default-Werten abweichen. Werden die Ergebnisse betrachtet, ist zu erkennen, dass bei beiden Datensätzen in unterschiedlichen Folds unterschiedliche Hyperparameterkombinationen ausgewählt wurden, was bei Random Forests ein häufiges Phänomen ist, da im Vergleich zu anderen Machine Learning Modellen der Einfluss der Hyperparameter geringer ist (Probst et al., 2019). Da sich die PRAUC-Werte über die verschiedenen Hyperparameterkombinationen in den Folds hinweg nur gering unterscheiden (Tabelle 4 und 6), kann davon ausgegangen werden, dass keine Kombination überlegen ist. Es lassen sich Tendenzen über die Folds hinweg erkennen und ableiten. So ist beim Hyperparameter der bei jedem Split zufällig ausgewählten Variablen zu erkennen, dass der häufig als Empfehlung diskutierte Wert $\sqrt{P}$ bei den meisten Folds deutlich überschritten wird, was darauf hindeutet, dass sich einige Variablen besser zur Prognose eignen als andere, was auch durch die Variablenwichtigkeit bestätigt wird (besonders bei der aktuellen Symptomatik zu beobachten; siehe Tabellen 3 und 5; Scornet, 2017). Überraschend ist, dass sich bei der Entscheidungsregel Hellinger nur selten gegen die anderen Entscheidungsregeln Extratrees und Gini durchsetzen konnte, obwohl diese bei unausgewogenen Daten Vorteile haben soll (Valavi et al., 2021). Eine mögliche Erklärung ist dabei, dass Hellinger für stark unabalancierte Daten entwickelt wurde und die anderen Hyperparameter auch einen Einfluss untereinander auf die Auswahl haben (Geurts et al.,

2006; Su, Ju, Liu & Yu, 2015). Gerade beim Datensatz Schweizer Haushalt-Panel wurde oft Gini als Entscheidungsregel ausgewählt, was auch erklären kann, wieso kategoriale Variablen nur eine geringe Variablenwichtigkeit haben, da Gini metrische Variablen präferiert (Provost, 2000; Valavi et al., 2021). Beim letzten betrachteten Hyperparameter geht es um die mindeste Anzahl Fälle in einem Knoten und somit um die Tiefe der Entscheidungsbäume. Hier wurde die grösste Spannbreite in den optimierten Hyperparametern gefunden, wobei der minimale Wert bereits so gross gewählt wurde, dass im Vergleich zum Defaultwert eins weniger Gefahr von Overfitting besteht (Barreñada et al., 2024; Bartz, 2023; Probst et al., 2019). Dabei ist zu erkennen, dass der tiefste übergebene Wert fünf nur zweimal ausgewählt wurde über alle Folds und Datensätze hinweg. Dies spricht dafür, dass höhere Werte zu einem besseren PRAUC-Wert bei den Testdaten der inneren Schleife führen. Eine mögliche Erklärung dafür ist, dass es ansonsten auf Grund der ausgewählten grossen Anzahl zufälliger Variablen pro Split zu Overfitting kommen könnte, wenn die Entscheidungsbäume zu tief werden würden (Barreñada et al., 2024; Probst et al., 2019). Die verbleibende Varianz bei ähnlichem PRAUC kann wiederum damit erklärt werden, dass die Hyperparameter bei Random Forests einen geringen Einfluss auf die Prognosegüte haben und sich gegenseitig beeinflussen (Probst et al., 2019). So kann beispielsweise die Entscheidungsregel Extratrees bei tiefen Bäumen die Testdaten besser vorhersagen als Gini, da wegen den zufällig gewählten Splits weniger Gefahr von Overfitting besteht (Geurts et al., 2006). Eine weitere Erklärung wäre, dass die PRAUC-Schätzungen bei 85 positiven Fällen beim Datensatz PSYCHE-D und bei 46 positiven Fällen beim Datensatz Schweizer Haushalt-Panel pro innerem Fold streuen können (Boyd, Eng & Page, 2013). Gerade beim Datensatz Schweizer Haushalt-Panel hätten demzufolge drei innere Folds zu einer stabileren Schätzung der PRAUC-Werte geführt. Dabei wären jedoch weniger Fälle zur Schätzung der Modelle zur Verfügung gestanden, was die Stabilität dieser bei der Anzahl Variablen beeinträchtigt hätte (Boyd et al., 2013; van der Ploeg et al., 2014). Eine Erhöhung der Anzahl äusseren Folds (beispielsweise auf zehn; Hastie et al., 2017) hätte nur

zu geringfügig mehr positiven Fällen pro innerem Fold geführt und es wurde auf Grund des erhöhten Rechenaufwands davon abgesehen.

Zuletzt wird zur Bewertung der Ergebnisse noch auf die lineare Diskriminanzanalyse eingegangen. Dabei ist zu erkennen, dass das Geschlecht, weniger jedoch das Alter, einen Einfluss auf die Schätzungen in den Marker- und Prävalenzmodellen hat. Somit ist die a-priori Wahrscheinlichkeit stärker vom Geschlecht und der generellen Prävalenz abhängig und weniger vom Alter der Person (Ramayah et al., 2010). Dies kann damit erklärt werden, dass der Effekt des Alters auf die depressive Symptomatik kurvilinear ist (Bundesamt für Statistik, 2022; Dao, Nguyen, Phung & Venkatesh, 2014). Bei den Geschlechtern zeigt sich, wie gemäss Literatur zu erwarten ist, ein höherer Wert für Frauen als für Männer (Eid, Gobinath & Galea, 2019; Hopcroft & Bradley, 2007; Weissman et al., 1993; WHO, 2017). Die Residuen der beiden Messzeitpunkte kovariieren, was einen Zusammenhang zeigt, jedoch nicht so hoch, dass die Abgrenzung der beiden Messzeitpunkte in Frage gestellt werden muss.

#### 5.1.2. Beantwortung erste Forschungsfrage zur Prognosegüte und Frühzeitigkeit.

Die erste in dieser Arbeit beschriebene Forschungsfrage befasst sich damit, ob eine adaptive, sequentielle Auswahl von longitudinalen Messsequenzen (online messbare Variablen) die mentale Gesundheit präzise und frühzeitig vorhersagen kann. Es konnte gefunden werden, dass mit einer sequentiellen Auswahl von Messzeitpunkten eine gute Prognosegüte erreicht werden kann, mit einer deutlichen Reduktion der Anzahl zweiter Messungen. Zwar gibt es bei der Beurteilung der BA keine klaren Normen, jedoch ist der in dieser Arbeit erreichte Wert deutlich über dem, der bei einer zufälligen Zuteilung zu erwarten gewesen wäre ($BA_{zufällig}$ = .5; Brodersen et al., 2010). Somit konnte die mentale Gesundheit präzise vorhergesagt werden. Bezüglich Frühzeitigkeit muss jedoch kritisch angemerkt werden, dass die aktuelle depressive Symptomatik allein zu einer gleich guten Prädiktionsleistung geführt hat wie die Prognose mit zusätzlichen Variablen, was die Gefahr von Zirkularität zeigt und somit eher das Beibehalten statt Entwickeln depressiver Symptome im kritischen Bereich vorhergesagt wird

(Chancellor et al., 2023). Jedoch konnte im Anhang A und B auch die Entwicklung depressiver Symptome über Threshold vorhergesagt werden ($BA_{\text{Maximal beim Datensatz PSYCHE-D}}$ = .74, $BA_{\text{Maximal beim Datensatz Schweizer Haushalt-Panel}}$ = .66), wobei auch je nach Kontext Vorteile der sequentiellen Auswahl der Messzeitpunkte gezeigt werden konnten. Dabei gilt bei diesen Modellen zu beachten, dass die Anzahl positiver Fälle gering ist, was die Stabilität der Ergebnisse einschränkt. Bei den Modellen in Anhang A und B, welche die Entwicklung der erhöhten depressiven Symptomatik vorhersagen, ist ausserdem zu erkennen, dass die anderen durch Boruta ausgewählten Prädiktionsvariablen neben der aktuellen depressiven Symptomatik im Vergleich eine höhere Variablenwichtigkeit aufweisen als bei den Modellen, welche das künftige Vorhandensein vorhersagen. Dies könnte ein Hinweis darauf sein, dass zusätzliche Variablen gegenüber der aktuellen Symptomatik eine grössere Rolle spielen bei der Erklärung der dynamischen Veränderung der Symptomatik als bei zukünftigen Zustandsprognosen. Als Grundlage dienen dabei die Variablen aus den im Theorieteil betrachteten Kategorien. Dies wäre in künftiger Forschung zu prüfen.

In dieser Arbeit konnte gezeigt werden, dass mit sequentieller Auswahl eine präzisere Vorhersage erreicht werden kann als mit einer einzelnen Messung. Dabei kann die Anzahl zweiter Messzeitpunkte deutlich reduziert werden im Vergleich zur starren Auswahl von zwei Messzeitpunkten. Somit kann in vielen Fällen durch sequentielle Auswahl eine frühzeitigere präzise Vorhersage erreicht werden als mit zwei Messzeitpunkten (wobei bezüglich Frühzeitigkeit die Frage nach der Zirkularität beachtet werden muss). Dies zeigt auf, dass die Ressourcen Zeit und Kosten bei der sequentiellen Auswahl auf die schwierigen Fälle verteilt werden können und geht mit Befunden aus früheren Arbeiten einher (Devine et al., 2016; Loe et al., 2017; Tan et al., 2018; Wong et al., 2025). Dabei ist jedoch neu bei der vorliegenden Arbeit, dass nicht die sequentielle Auswahl von Items und Messmethoden untersucht wurde, sondern die sequentielle Auswahl von Messzeitpunkten bei der Vorhersage der depressiven Symptomatik.

Somit kann mit Blick auf die vorhergesagte zukünftige depressive Symptomatik im kritischen Bereich anhand der aktuellen Ausprägung aus praktischer Sicht entschieden werden, ob ein Eingreifen nötig ist, eine zweite Messung abgewartet wird oder kein Eingreifen nötig ist, da in Zukunft die depressive Symptomatik den definierten Threshold nicht übersteigt. Dies ermöglicht ein frühzeitigeres Eingreifen bei gleicher Genauigkeit, als wenn immer ein zweiter Messzeitpunkt abgewartet wird und eine höhere Genauigkeit, als wenn immer nur ein Messzeitpunkt verwendet wird. Dabei stellt sich gerade beim Datensatz PSYCHE-D die Frage nach dem Zeitabstand zwischen den Messungen (siehe Abbildung 3 und 4 und Anhang O). Diese Vorteile können auch genutzt und gefunden werden, wenn bloss die aktuelle depressive Symptomatik zur Prognose verwendet wird (Anhang M). So können beispielsweise interventionelle Massnahmen direkt angewendet werden bei Personen, die bereits nach einer Messung klar positiv zugeteilt werden können und bei Personen, welche nicht klar zugeteilt werden können, eine weitere Messung abgewartet werden. So könnten therapeutische Ressourcen sparsam eingesetzt werden, ohne dabei die Prognosefähigkeit von zwei Messzeitpunkten zu beeinträchtigen (Abbildungen 2 und 3).

**5.1.3. Beantwortung zweite Forschungsfrage zu Kosten.** Die zweite Forschungsfrage beinhaltet, ob mit einer sequentiellen Auswahl von Messungen die mentale Gesundheit kosteneffizienter vorhergesagt werden kann als mit einer starren Auswahl von Messungen, die gleich für alle Personen ist. Dies ist ein wichtiges Ziel des Multi-Objective Ansatzes, da neben der Genauigkeit und Frühzeitigkeit auch die Kosten eine wichtige Rolle spielen (Roijers et al., 2013). Es konnte gefunden werden, dass im Vergleich zu zwei starren Messzeitpunkten die sequentielle Auswahl von Messzeitpunkten bei allen definierten Messkosten weniger durchschnittliche Gesamtkosten verursachte. Währenddem der Unterschied bei tiefen Messkosten noch gering ist, zeigen sich grosse Unterschiede und Vorteile bei steigenden definierten Messkosten. Beim Vergleich von einem starren Messzeitpunkt mit der sequentiellen Auswahl wurden weniger deutliche Ergebnisse und

Kostenvorteile der sequentiellen Auswahl gefunden. Verursacht die sequentielle Auswahl bei allen vier Vergleichen bei tiefen definierten Messkosten geringere durchschnittliche Gesamtkosten als eine starre Messung, nähern sich die durchschnittlichen Gesamtkosten bei steigenden Messkosten aneinander an. Bei gewissen Kostenkombinationen wurde sogar gefunden, dass die durchschnittlichen Messkosten der sequentiellen Auswahl die der starren einmaligen Messung gering übersteigen. Dies ist insbesondere bei den Modellen zu beobachten, bei denen die Unterschiede in der Prognosegüte von einer starren und zwei starren Messzeitpunkten nur gering sind, was gerade beim Datensatz PSYCHE-D auf den geringen Zeitabstand von drei Monaten zwischen den Messungen zurückgeführt werden könnte. Dabei ist zu beobachten, dass insbesondere die Sensitivität beim Datensatz PSYCHE-D nicht gesteigert wird durch die zweite Messung bei der Gewichtung nach inversem Anteil und somit nur wenig Möglichkeit für die sequentielle Auswahl besteht, die höher gewichteten falsch negativ Kosten zu reduzieren (auch anhand der PRAUC-Werte in Tabelle 6 zu erkennen). Dies könnte eine Erklärung für die höheren durchschnittlichen Gesamtkosten sein bei steigenden definierten Messkosten.

Bezüglich dem betrachteten Zeitabstand wäre es spannend systematisch zu prüfen, inwiefern die Veränderung der Zeitspanne zwischen den Messungen einen Einfluss auf die Ergebnisse hat. So wäre es denkbar, dass bei einer grösser gewählten Zeitspanne die Prognosegüte zwischen einer starren Messung und zwei starren Messungen weiter auseinandergehen und die sequentielle Auswahl somit auch bezüglich Kosten mehr Vorteile gegenüber einer starren Messung entfalten kann (erste Hinweise im Anhang O). Dies wäre aus praktischer Sicht relevant für die Entscheidung, in welchen Zeitabständen regelmässige Monitorings und Screenings durchgeführt werden sollten, um die gemäss Multi-Objective Ansatz wichtigen Aspekte Frühzeitigkeit, Kosten und Genauigkeit zu optimieren (Roijers et al., 2013). Die Betrachtung des Zeitabstands ist weiterführend von Bedeutung, da die US Preventive Services Task Force (2023) Screening für depressive Störungen empfiehlt, jedoch

auch berichtet, dass noch nicht klar ist, in welchem Zeitabstand dies gemacht werden sollte und was die potentiellen Nachteile und Risiken eines zu regelmässigen Screenings sind.

Folglich lassen sich besonders bei tiefen definierten Messkosten Vorteile der sequentiellen Auswahl im Vergleich zur einmaligen starren Messung identifizieren. Dabei kann aus praktischer Sicht argumentiert werden, dass es sich beim Datensatz PSYCHE-D grossenteils um passiv gemessene Daten handelt und somit die Messkosten für die Personen gering sind (Liang et al., 2019). Auch durch die automatisierte Erhebung und Auswertung mittels Algorithmen können die Kosten realistisch betrachtet gering gehalten werden (Liang et al., 2019; Onnela, 2021). Beim Datensatz Schweizer Haushalt-Panel wurden zur Prädiktion 13 selbstberichtete Fragen verwendet. Der Aufwand für die Personen ist somit grösser einzuschätzen als bei den passiv gemessenen Daten beim Datensatz PSYCHE-D. Trotzdem haben selbstberichtete Fragen den Vorteil, wenig Zeit in Anspruch zu nehmen und können selbstständig beantwortet werden (Reinders, 2011). Ausserdem kann die Auswertung automatisiert und selbstständig erfolgen, was für geringe Messkosten spricht (Reinders, 2011). Zuletzt handelt es sich bei allen Messungen um nicht-invasive Messungen, was auch physische Kosten der Beteiligten geringhält. Dabei ist jedoch zu beachten, dass alle Fragen beantwortet werden müssen für eine präzise Prädiktion, was die Qualität der Antworten negativ beeinflussen und den Dropout erhöhen kann (Décieux, Mergener, Neufang & Sischka, 2015). Dabei wurde in dieser Arbeit darauf geachtet, dass eine geringe Anzahl an Variablen ausgewählt wird, da jeweils nur Variablen betrachtet wurden, welche zu beiden Messzeitpunkten als relevant bewertet wurden. Folglich kann argumentiert werden, dass die definierten Messkosten in der vorliegenden Arbeit aus praktischer Sicht als gering beurteilt und gewählt werden können. In diesen Kontexten verursacht die sequentielle Auswahl gegenüber der starren Auswahl von einem oder zwei Messzeitpunkten geringere durchschnittliche Gesamtkosten. Dabei muss erwähnt werden, dass die zusätzlichen Variablen zu keiner besseren Prädiktionsleistung gemessen an der BA geführt haben als die blosse

Betrachtung der depressiven Symptomatik. Folglich wäre es am kosteneffizientesten, zur Vorhersage die sequentielle Auswahl von Messzeitpunkten der depressiven Symptomatik zu verwenden. Die Anwendung der sequentiellen Algorithmen führte dabei auch bei der Prädiktion nur mit der aktuellen depressiven Symptomatik zu bestätigenden Ergebnissen, was die relative Prognosegüte, gemessen an der BA, und die vergleichende Kostenstruktur bei einem, zwei und sequentiell ausgewählten Messzeitpunkten betrifft (Proof of Concept; Anhang M sowie Abbildungen 2 und 3).

Mit Blick auf die Kosten muss auch genannt werden, dass diese abhängig vom Kontext sind. Stehen beispielsweise nur wenige Wearables zur Verfügung, kann es sinnvoll sein, die Messkosten im Verhältnis zu den Fehlklassifikationskosten höher zu wählen, damit eine geringere Anzahl zweiter Messzeitpunkte ausgewählt wird. Dabei gilt bezüglich der Messkosten auch zu beachten, dass bei der Auswahl einer zweiten Messung eine längere Zeit abgewartet werden muss (in dieser Arbeit beim Datensatz Schweizer Haushalt-Panel ein Jahr und beim Datensatz PSYCHE-D drei Monate), bevor eine Zuteilung vorgenommen wird und mit der Intervention begonnen werden kann, was sich negativ auf den Behandlungserfolg auswirken kann (Kraus et al., 2019). Eine Option wäre in der Zeit, in der die zweite Messung abgewartet wird, mit niedergeschwlligen und kostengünstigen Massnahmen zu beginnen (beispielsweise mit kostenlosen Applikationen, die metaanalytisch nachgewiesen depressive Symtome reduzieren können; Linardon, Messer, Goldberg & Fuller-Tyszkiewicz, 2023). Weitergehend ergibt sich der Vorteil, dass durch das Abwarten der zweiten Messung die Prognosegenauigkeit steigt und somit weniger Personen fälschlicherweise oder garnicht Massnahmen erhalten. Auch die Fehlklassifikationskosten können an den Kontext angepasst werden. So kann es bei wenig vorhandenen therapeutischen Ressourcen sinnvoll sein, die gleiche Gewichtung der Fehlklassifikationskosten zu wählen, um nur die deutlichen positiven Fälle zu identifizieren (Brehaut et al., 2023; US Preventive Services Task Force, 2002). Steht hingegen mehr im Fokus, dass jede Person mit künftigen starken depressiven Symptomen

identifiziert werden soll und es weniger schlimm ist, wenn Personen fälschlicherweise Interventionen erhalten, empfiehlt sich die inverse Gewichtung nach Anteil (Brehaut et al., 2023). Weiterführend können beispielsweise bei grosser Wichtigkeit einer frühzeitigen Prognose die Messkosten höher definiert werden.

Diese Steigerung der Prognosegüte mit einer Verringerung der Messkosten und kürzeren Beobachtungszeiten entspricht dem Multi-Objective Ansatz (Roijers et al., 2013). In dieser Arbeit konnte gezeigt werden, dass ergänzend zu den bereits etablierten Befunden der sequentiellen Auswahl von Messmethoden auch die sequentielle Auswahl von Messzeitpunkten derselben Messmethoden die verschiedenen Aspekte des Multi-Objective Ansatz optimiert und ausbalanciert werden können (Devine et al., 2016; Loe, et al., 2017; Tan et al., 2018; Wong et al., 2025). So ermöglicht dieser Ansatz aus praktischer Sicht beispielsweise eine adaptive Gestaltung von einem regelmässigen Monitoring in sequentieller Form. Bei einer klaren Zuteilung von künftig geringer depressiver Symptomatik könnte über einen längeren Zeitraum auf weitere Erhebungen verzichtet werden, während bei unsicherer Zuteilung und somit höherem Risiko eine frühere Erhebung erforderlich wäre (Nease & Malouin, 2003). Gleichzeitig könnten bei identifizierten Risikofällen frühzeitig geeignete Interventionen eingeleitet werden.

## 5.2. Limitationen

Auch diese Arbeit hat Limitationen, die im Folgenden genannt und diskutiert werden. Die erste Limitation gilt den untersuchten Variablen. So wurde die abhängige Variable beim Datensatz Schweizer Haushalt-Panel nicht über eine validierte klinisch-psychologische Messung geformt, sondern über die vorhandenen Variablen gebildet, welche vermutlich nur einen Aspekt der depressiven Symptomatik erfassen (Watson et al., 1988; WHO, 2024). Dies zeigt eine Gefahr vom datenbasierten Vorgehen mit Big Data auf (Adjerid & Kelley, 2018). Dies führte bei den Daten, bei denen erfasst wurde, ob eine depressive Störung diagnostiziert

wurde, zu einer Sensitivität von 66.2% und einer Spezifität von 87.1%. Beim Datensatz PSYCHE-D konnte für die Operationalisierung der abhängigen Variable der validierte Fragebogen PHQ-9 und Threshold grösser zehn verwendet werden, jedoch führte dies zu einem grösseren Anteil an positiven Fällen als in der Allgemeinbevölkerung (Bundesamt für Statistik, 2022; Kroenke et al., 2001; Levis et al., 2019; Manea et al., 2012; Moriarty et al., 2015). Hinzu kommt, dass für die Zuteilung nur ein Zeitpunkt betrachtet wurde, was problematisch sein kann, da die Ausprägung der Symptome variieren kann und unter anderem von der Interpretation der Fragen abhängt (Kokanovic et al., 2013; Musliner et al., 2016; Ong et al., 2022). Auch bei den Prädiktionsvariablen konnten psychologische Konstrukte nur wenig berücksichtigt werden, da diese im PSYCHE-D gar nicht und im Schweizer Haushalt-Panel mit einem zu grossen Zeitabstand erfasst wurden, wobei die Variablen nicht aus validierten Fragebögen stammen. Eine Durchführung mit psychosozialen Variablen und dreijährigem Abstand zwischen dem ersten und zweiten Messzeitpunkt sowie eine Beschreibung der verwendeten Variablen ist in Anhang Q zu sehen, wobei unter anderem auf Glaubenssätze zum Kontrollgefühl eingegangen wurde. Auch dort ist zu erkennen, dass die aktuelle depressive Symptomatik die höchste Variablenwichtigkeit aufweist. Es wäre darüber hinaus spannend gewesen zu prüfen, ob mit psychologisch validierten Variablen und Konzepten die Prognosegüte verbessert worden wäre, da das Diathese-Stress-Modell davon ausgeht, dass die Vulnerabilität beispielsweise über die Persönlichkeit erfasst werden kann (Colodro-Conde et al., 2018; Flett et al., 1995). Ein weiteres Beispiel wäre die validierte Messung des wahrgenommenen Stresses mit der Perceived Stress Scale (Klein et al., 2016). Eine Möglichkeit dies aufzunehmen, wäre in Zukunft die regelmässige mehrmalige Messung dieser Variablen für die zeitlich sequentielle Auswahl oder die Berücksichtigung im Prävalenzmodell in der Diskriminanzanalyse von zeitlich stabileren Konstrukten wie der Persönlichkeit. Bei Letzterem stellt sich jedoch die Frage nach der Stabilität dieser Messungen (Caspi, Roberts & Shiner, 2005). Bei den biometrischen Messungen ist als

Limitation zu nennen, dass nur die Schritt- und Schlafdaten aus den Messungen mit Wearables zur Verfügung standen. Dabei konnte beispielsweise bereits gezeigt werden, dass die Herzratenvariabilität ein wichtiger biometrischer Prädikator für die mentale Gesundheit ist (Perna et al., 2020; Ramesh, Nayak, Beestrum, Quer & Pandit, 2023).

Eine weitere Limitation der Arbeit auf Grund der verwendeten Daten ist, dass nicht vorhergesagt werden konnte, welche Personen eine depressive Störung entwickeln, sondern stattdessen untersucht wurde, welche Personen zu einem späteren Zeitpunkt erhöhte depressive Symptome haben. Dabei wäre die Vorhersage der Entwicklung einer depressiven Störung besonders mit Blick auf die wichtige frühzeitige Intervention von Bedeutung (Kraus et al., 2019; Nazareth & Kendrick, 2014). Dies konnte jedoch nicht als Hauptuntersuchung durchgeführt werden, da dieses Vorgehen zu einer zu geringen Anzahl positiver Fälle führte, was instabile Ergebnisse zur Folge hat (siehe Anhang R). Der Vollständigkeit halber wurden die beschriebenen Analysen mit Stables und Converters auch durchgeführt und die Ergebnisse sind im Anhang A und B zu finden. Diese Anpassung führt zur Konfundierung, ob die ausgewählten Messungen und erstellten Modelle die Aufrechterhaltung oder die Entwicklung von depressiver Symptomatik im kritischen Bereich vorhersagen. Dies führt ausserdem dazu, dass die aktuelle depressive Symptomatik einen grossen Einfluss auf die Schätzung hat (siehe Tabelle 3 und 5) und die erstellten Modelle zu keiner besseren BA führen als die Prognose nur mit der aktuellen depressiven Symptomatik. Dies verweist auf die Gefahr von Zirkularität (Chancellor et al., 2023).

Eine weitere Limitation, welche sich auf die biometrischen Messungen bezieht, ist, dass sich die physiologischen Reaktionen auf Stress und dessen Auswirkungen zwischen Menschen stark unterscheiden (Glass, Lake, Contrada, Kehoe & Erlanger, 1983; Jacoby et al., 2021). Folglich ist es schwierig, Indikatoren ausfindig zu machen, welche allgemeine Gültigkeit haben. Gerade bei längsschnittlichen Daten wäre es demzufolge sinnvoll, mit Multilevel Modellen zu arbeiten, bei denen feste und variable Effekte pro Person erfasst

werden können (Peugh, 2010). In der vorliegenden Arbeit war dies jedoch nur eingeschränkt möglich, da pro Person nur zwei Messzeitpunkte der unabhängigen Variablen betrachtet wurden. Somit war die Modellierung mit einem Random Slope nicht möglich. Künftige Forschung könnte hier entweder mehr als zwei Prognosezeitpunkte verwenden oder in einer ersten Phase personenspezifische Modellierungen erfassen und darauffolgend eine zweite Phase mit sequentiell ausgewählten Messzeitpunkten durchführen. Dabei wurden beim Datensatz PSYCHE-D auch Variablen verwendet, welche die Veränderung im Verhaltensmuster erfassten. Diese reichten jedoch nur maximal vierzehn Tage zurück.

Als Limitationen der statistischen und methodischen Vorgehensweise gelten die fehlende Datenimputation, die Variablenauswahl, welche nicht mit nested Cross Validation gemacht wurde und die Erfassung des Alters im Prävalenzmodell. Bei der fehlenden Datenimputation liegt das Problem darin, dass sowohl beim Datensatz PSYCHE-D als auch beim Datensatz Schweizer Haushalt-Panel viele fehlende Messungen vorliegen. Einerseits liegt das beim Schweizer Haushalt-Panel daran, dass gewisse Variablen nicht jedes Jahr erhoben werden und gewisse Folgefragen nur an Personen gerichtet werden, welche eine vorhergehende Frage entsprechend beantwortet hatten. Dies konnte
gelöst werden, indem nur Variablen mit einem geringen Anteil fehlender Werte berücksichtigt wurden. Auf der anderen Seite liegt das bei beiden Datensätzen daran, dass einige Messungen nicht vollständig sind. Werden die Kapitel 3.1. und 3.2. dieser Arbeit betrachtet, ist zu erkennen, dass es sich unter anderem um systematische Stichprobenausfälle handelt, da der Anteil an Personen mit depressiven Störungen durch Entfernen der fehlenden Werte reduziert wird. Dies ist problematisch, da systematische Stichprobenausfälle zu verzerrten Ergebnissen, weniger Generalisierbarkeit und Unter- sowie Überschätzung der Effekte führen können (Blickle, 2019; Schupp & Wolf, 2015). Eine mögliche Lösung ist die Datenimputation, wovon jedoch bei dieser Arbeit auf Grund des hohen Aufwands abgesehen wurde (Jäger, Allhorn & Biessmann, 2021). Besonders aufwändig ist die Imputation der Daten, da komplexere

Modelle, welche beispielsweise mit Random Forests erzeugt werden, besser abschneiden als welche, die einfacher sind, was besonders bei den systematischen Stichprobenausfällen eine Herausforderung darstellt (Jäger et al., 2021). Eine weitere Möglichkeit mit fehlenden Daten umzugehen, ist mit Algorithmen wie XGBoost zu arbeiten, welche flexibler mit fehlenden Werten umgehen können und diese in die Modellierung aufnehmen (Fatima, Hussain, Amir, Ahmed & Aslam, 2023).

Bei der Limitation der Variablenauswahl ohne nested Cross Validation besteht die Gefahr von Overfitting (Reunanen, 2003). Da die Variablen durch dieselben Daten ausgewählt wurden, auf die sie später auch angewendet wurden, besteht die Gefahr, dass sich die ausgewählten Variablen beim aktuellen Datensatz zur Prognose eignen, diese jedoch nicht auf andere Daten angewendet werden können (Dietterich, 1995). Dies wurde gemacht, da eine Variablenauswahl nach Fold zu aufwendig gewesen wäre und die Interpretierbarkeit der Ergebnisse erschwert hätte. Ausserdem hätte die geringere Anzahl an Werten bei den vielen Variablen die Stabilität der Ergebnisse reduziert (Kursa & Rudnicki, 2010). Dabei ist eine weitere Limitation für die Reproduzierbarkeit, dass beim Datensatz PSYCHE-D kein Seed gesetzt wurde bei der Variablenauswahl mit Boruta. Beim Datensatz Schweizer Haushalt-Panel wurde dies gemacht. Zur Kontrolle der Stabilität wurden zu einem späteren Zeitpunkt verschiedene Seeds ausprobiert und die Variablen, die jeweils ausgewählt wurden, decken sich mit denen, die in dieser Arbeit zur Prädiktion verwendet wurden, wobei eine hohe Stabilität der Auswahl zu beobachten ist (es wurden 15 Seeds ausprobiert und die Häufigkeit der ausgewählten Variablen ist im Anhang S aufgeführt). Besonders deutlich ist dies bezüglich ausgewählter Kategorien von Variablen zu erkennen und Variablen, welche gemäss Tabelle 3 und 5 eine hohe Variablenwichtigkeit aufweisen.

Als weitere Limitation des methodischen und statistischen Vorgehens wird die Modellierung des Effekts des Alters in der Diskriminanzanalyse in den Prävalenzmodellen und den Markermodellen genannt. Der Effekt des Alters auf die depressive Symptomatik

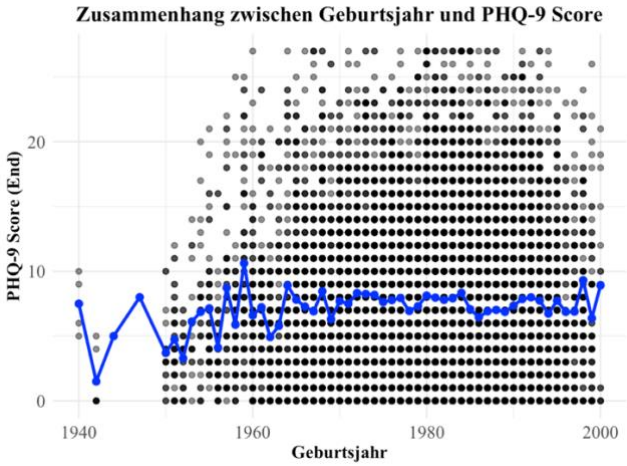


*Abbildung 7.* Zusammenhang zwischen Geburtsjahr und PHQ-9 Wert PSYCHE-D.
*Anmerkung.* Die blaue Linie entspricht dem mittleren PHQ-9 Score für das Geburtsjahr und somit des aktuellen Alters zum Zeitpunkt der Messung.

wurde als linear erfasst. Hier wäre es sinnvoller gewesen, das Alter in kategorialen Einteilungen in Gruppen zu erfassen, den Effekt als nicht linear zu modellieren oder anderen Verfahren, wie beispielsweise Random Forests, für die Schätzung des Prävalenzmodells zu verwenden. Jedoch hat sich, wie in Abbildung 7 zu erkennen ist, exemplarisch für den Datensatz PSYCHE-D gezeigt, dass auch bezüglich Altersgruppen nur geringe Unterschiede bezüglich PHQ-9 Mittelwert zu erkennen sind, da die Abweichungen innerhalb der Altersgruppen gross sind. Auch auf den gesamten Datensatz PSYCHE-D angewendet, konnte keine signifikante Korrelation zwischen Alter und PHQ-9-Wert gefunden werden ($r = .02$, $p = .08$).

Als technischer Fehler muss erwähnt werden, dass irrtümlicherweise die Variable «Fold» den Random Forest Modellen beim Datensatz PSYCHE-D zur Prädiktion übergeben wurde. Dabei erreichte diese nur eine geringe Variablenwichtigkeit von $M = 3.99$ ($SD = 3.58$). Dies zeigt einen geringen Einfluss auf die Modellleistung und ein systematischer Einfluss auf die Schätzung kann ausgeschlossen werden, da die finalen Schätzungen jeweils auf Modellen beruhen, die die Daten nicht zur Erstellung der Random Forests verwendet haben und somit der Fold identisch ist. Eine stärkere Problematik hätte sich ergeben, wenn der Anteil an Personen mit künftiger erhöhter depressiver Symptomatik in den Folds unterschiedlich

gewesen wäre, was bei der vorliegenden Arbeit jedoch nicht der Fall ist. Zur Kontrolle wurde mit denselben Daten wie in der Hauptuntersuchung der Arbeit ein Modell erstellt, ohne die Variable «Fold» zur Prädiktion. Dies führte bei der Prognose mit einem festen und zwei festen Messzeitpunkten zur selben auf zwei Nachkommastellen gerundeten BA bei Cutoff .5 und Cutoff Prävalenz.

Als Limitation der sequentiellen Entscheidungsregeln gilt die Subjektivität der definierten Kosten der Messung und der Fehlklassifikation (Wyss et al., 2023). Dabei wird argumentiert, dass die praktische Anwendung bislang wegen fehlendem Vertrauen in die Entscheidungsregeln erst selten ist (Denton, 2018).

Zuletzt muss noch erwähnt werden, dass die Daten für die Untersuchung jeweils nur aus einem Kulturkreis stammen. So wurden beim Datensatz PSYCHE-D nur Personen aus den USA untersucht und beim Datensatz Schweizer Haushalt-Panel nur Personen aus der Schweiz. Es ist fraglich, ob sich die Ergebnisse und identifizierten Variablen auch auf andere Kulturkreise anwenden lassen (Stewart et al., 2004). Ausserdem wurden beim Datensatz Schweizer Haushalt-Panel nur Personen berücksichtigt, die angegeben haben, arbeitstätig zu sein, was die Generalisierbarkeit der Ergebnisse vermindert. Dabei ist auch zu berücksichtigen, dass der Zeitpunkt des Jahres der Befragung, welcher sich beim Datensatz Schweizer Haushalt-Panel zwischen den Personen und beim Datensatz PSYCHE-D innerhalb der Personen unterscheidet, einen Einfluss auf die abhängige Variable haben kann (Lukmanji, Williams, Bulloch, Bhattarai & Patten, 2019; Lukmanji, Williams, Bulloch & Patten, 2020). Dies wurde nicht modelliert. Dabei ist beim Datensatz PSYCHE-D auch fraglich, welchen Effekt das Tragen der Wearables und das Wissen über die Datenerfassung auf die Aktivität und das Verhalten hat, was zu einer Verfälschung der Ergebnisse führen kann (Brickwood, Watson, O`Brien & Williams, 2019).

## 5.3. Fazit und Ausblick

In dieser Arbeit konnte im Rahmen des Multi-Objective Ansatzes gezeigt werden, dass mit der sequentiellen Auswahl von Messzeitpunkten derselben Variablen eine Balance aus Kosten, Präzision und Frühzeitigkeit der Prognose erreicht werden kann zur Prädiktion der zukünftigen erhöhten depressiven Symptomatik (Roijers et al., 2013). Dabei wurden zum einen Variablen aus dem Arbeitsleben, zum Lebensstil, zur Gesundheit und zur Freizeit verwendet. Zum anderen wurden Variablen von biometrischen Messungen verwendet. Beide Variablenkategorien konnten zusammen mit der aktuellen depressiven Symptomatik mittels Random Forrests das zukünftige Vorhandensein depressiver Symptomatik im kritischen Bereich nicht besser vorhersagen als die blosse Erfassung der aktuellen depressiven Symptomatik zum Prognosezeitpunkt. Im Vergleich zu zwei Messzeitpunkten verursachten die sequentiell ausgewählten Messzeitpunkte durchgehend geringere durchschnittliche Kosten. Der Unterschied wurde bei steigenden definierten Messkosten deutlich grösser. Im Vergleich zu der starren Auswahl von einem Messzeitpunkt hingegen konnte die sequentielle Auswahl die durchschnittlichen Kosten nur bei geringen definierten Messkosten reduzieren, wobei teilweise bei steigenden Messkosten die durchschnittlichen Gesamtkosten der sequentiellen Auswahl etwas höher waren als die der starr ausgewählten einzelnen Messung. Der Vorteil der sequentiellen Auswahl, dass sowohl eine wie auch zwei starre Messzeitpunkte höhere durchschnittliche Gesamtkosten verursachen, zeigt sich folglich vor allem bei tiefen definierten Messkosten. Dies lässt sich teilweise dadurch erklären, dass die Prädiktionsleistung von einer einzigen Messung ähnlich hoch war wie der von der gemeinsamen Prognose mit zwei Messungen. Zukünftige Forschung sollte hier weitere Untersuchungen durchführen mit grösseren Zeitabständen. Ausserdem sollte sich künftige Forschung darauf fokussieren, die Entwicklung depressiver Symptome vorherzusagen statt wie in dieser Arbeit, wegen geringer Stichprobenzahl, bloss das zukünftige Vorhandensein der Symptomatik. Zuletzt wäre es interessant Prädiktionsmodelle zu entwickeln, welche

biometrische und selbstberichtete Angaben kombinieren sowie psychologisch validierte Messmethoden zur zeitlich sequentiellen Auswahl nutzen. Dabei können auch die vorliegenden Datensätze weiter analysiert und verschiedene Machine Learning Verfahren angewendet werden, um zu prüfen, ob inkrementelle Varianz im Vergleich zur aktuellen depressiven Symptomatik erklärt werden kann. Für die Praxis können die sequentiellen Entscheidungsregeln von grosser Bedeutung sein, da sie eine Möglichkeit bieten, die Messkosten zu reduzieren, ohne dabei die Genauigkeit der Prognose zu vermindern. Eine mögliche Anwendung ist ein regelmässiges, adaptives Monitoring, wobei die Kosten so definiert werden können, dass sich die Entscheidungsregeln an den Kontext anpassen.

# Literaturverzeichnis

Adjerid, I. & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899–917. https://doi.org/10.1037/amp0000190

Adler, D. A., McLaughlin, T. J., Rogers, W. H., Chang, H., Lapitsky, L. & Lerner, D. (2006). Job performance deficits due to depression. *American Journal of Psychiatry*, *163*(9), 1569– 1576. https://doi.org/10.1176/ajp.2006.163.9.1569

Aler, R., Valls, J. M., & Boström, H. (2020). Study of Hellinger Distance as a splitting metric for Random Forests in balanced and imbalanced classification datasets. *Expert Systems with Applications*, *149*, 113264. https://doi.org/10.1016/j.eswa.2020.113264

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues*, *9*(5), 272-278.

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress*, *40*(7), 64-65.

Andersson, S., Bathula, D. R., Iliadis, S. I., Walter, M., & Skalkidou, A. (2021). Predicting women with depressive symptoms postpartum with machine learning methods. *Scientific Reports*, *11*(1), 7877. https://doi.org/10.1038/s41598-021-86368-y

Andrews, B., Qian, M., & Valentine, J. D. (2002). Predicting depressive symptoms with a new measure of shame: The Experience of Shame Scale. *British Journal of Clinical Psychology*, *41*(1), 29-42. https://doi.org/10.1348/014466502163778

Andridge, R. R. & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40–64. https://doi.org/10.1111/j.1751- 5823.2010.00103.x

Angst, J., Angst, F. & Stassen, H. H. (1999). Suicide risk in patients with major depressive disorder. *The Journal of Clinical Psychiatry*, *60*(2), 57-62. https://pumed.ncbi.nlm.nih.gov/10073389/

Araf, I., Idri, A. & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: A review. *Artificial Intelligence Review*, *57*(4), 57-80. http://doi.org/10.1007/s10462-023-10652-8

Archer, T., Josefsson, T. & Lindwall, M. (2014). Effects of physical exercise on depressive symptoms and biomarkers in depression. *CNS & Neurological Disorders Drug Targets*, *13*(10), 1640-1653. https://doi.org/10.2174/1871527313666141130203245

Arnaldo, I., Corcoran, A. W., Friston, K. J., & Ramstead, M. J. (2022). Stress and its sequelae: An active inference account of the etiological pathway from allostatic overload to depression. *Neuroscience & Biobehavioral Reviews*, *135*, 104590. https://doi.org/10.1016/j.neubiorev.2022.104590

Asarnow, L. D. (2020). Depression and sleep: What has the treatment research revealed and could the HPA axis be a potential mechanism?. *Current Opinion in Psychology*, *34*, 112–116. https://doi.org/10.1016/j.copsyc.2019.12.002

Austin, R. R., McLane, T. M., Pieczkiewicz, D. S., Adam, T. & Monsen, K. A. (2023). Advantages and disadvantages of using theory-based versus data-driven models with social and behavioral determinants of health data. *Journal of the American Medical Informatics Association*, *30*(11), 1818–1825. https://doi.org/10.1093/jamia/ocad148

Balayla, J. (2024). Prevalence threshold and public health. In J. Balayla (Ed.), *Theorems on the prevalence threshold and the geometry of screening curves: A bayesian approach to clinical decision-making* (pp. 187-203). Cham: Springer. https://doi.org/10.1007/978-3-031-71452-8_14

Barreñada, L., Dhiman, P., Timmerman, D., Boulesteix, A.-L. & Van Calster, B. (2024). Understanding overfitting in random forest for probability estimation: A visualization and simulation study. *Diagnostic and Prognostic Research*, *8*(1), 14. https://doi.org/10.1186/s41512- 024-00177-1

Bartz, E. (2023). *Hyperparameter tuning for machine and deep learning with R: A practical guide.* (1st ed.). Singapore: Springer.

Becker, S. P., Langberg, J. M. & Evans, S. W. (2015). Sleep problems predict comorbid externalizing behaviors and depression in young adolescents with attention-deficit/hyperactivity disorder. *European Child & Adolescent Psychiatry*, *24*(8), 897–907. https://doi.org/10.1007/s00787-014-0636-6

Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., & West, M. (2007). Generative or discriminative? Getting the best of both worlds. *Bayesian Statistics*, *8*(3), 3-24.

Berrar, D. (2019). Cross-validation. In S. Ranganathan, M. Gribskov, K. Nakai & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (pp. 542–545). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2020). *Supervised and unsupervised learning for data science* (pp. 3–21). Cham: Springer.

Bertrand, M., & Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. *American Economic Review*, *91*(2), 67-72. https://doi.org/10.1257/aer.91.2.67

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Binnewies, J., Nawijn, L., van Tol, M.-J., van der Wee, N. J. A., Veltman, D. J. & Penninx, B. W. J. H. (2021). Associations between depression, lifestyle and brain structure: A longitudinal MRI study. *NeuroImage*, *231*, 117834. https://doi.org/10.1016/j.neuroimage.2021.117834

Blickle, G. (2019). Methoden. In F.W. Nerdinger, G. Blickle & N. Schaper (Hrsg.), *Arbeits- und Organisationspsychologie* (S. 29–44). Heidelberg: Springer. https://doi.org/10.1007/978-3-662-56666-4_3

Boumparis, N., Karyotaki, E., Kleiboer, A., Hofmann, S. G. & Cuijpers, P. (2016). The effect of psychotherapeutic interventions on positive and negative affect in depression: A systematic review and meta-analysis. *Journal of Affective Disorders*, *202*(1), 153–162. https://doi.org/10.1016/j.jad.2016.05.019

Boyd, K., Eng, K. H. & Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. In H. Blockeel, K. Kersting, S. Nijssen & F. Železný (Eds.), *Machine learning and knowledge discovery in databases* (pp. 451–466). Heidelberg: Springer. http://doi.org/10.1007/978-3-642-40994-3_55

Brehaut, E., Neupane, D., Levis, B., Wu, Y., Sun, Y., Ioannidis, J. P., ... & Thombs, B. D. (2023). 'Optimal'cutoff selection in studies of depression screening tool accuracy using the PHQ-9, EPDS, or HADS-D: A meta-research study. *International Journal of Methods in Psychiatric Research*, *32*(3), e1956. https://doi.org/10.1002/mpr.1956

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brickwood, K. J., Watson, G., O'Brien, J., & Williams, A. D. (2019). Consumer-based wearable activity trackers increase physical activity participation: Systematic review and meta-analysis. *JMIR mHealth and uHealth*, *7*(4), e11819. https://doi.org/10.2196/11819

Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *2010 20th International Conference on Pattern Recognition* (pp. 3121–3124). New Jersey: IEEE. https://doi.org/10.1109/ICPR.2010.764

Buchan, I. E. (2025). *Normal scores transform*. In StatsDirect. Zugriff am 02.03.2025. Abgerufen unter: https://www.statsdirect.com/help/data_preparation/transform_normal_scores.htm

Bundesamt für Statistik. (2022). *Psychische Störungen: beispielloser Anstieg der Hospitalisierungen bei den 10- bis 24-jährigen Frauen - Behandlung von psychischen Störungen bei jungen Menschen in den Jahren 2020 und 2021*. Zugriff am 18.09.2024. Abgerufen unter: https://www.bfs.admin.ch/asset/de/23772011

Bundesamt für Statistik. (2024). *Die Gesundheitskosten stiegen 2022 auf 91,5 Milliarden Franken*. Zugriff am 19.02.2025. Abgerufen unter: https://www.bfs.ad- min.ch/asset/de/31225971

Cacheda, F., Fernandez, D., Novoa, F. J. & Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *Journal of Medical Internet Research*, *21*(6), e12554. https://doi.org/10.2196/12554

Caspi, A., Roberts, B. W. & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, *56*(1), 453–484. https://doi.org/10.1146/annurev.psych.55.090902.141913

Chancellor, S., Feuston, J. L., & Chang, J. (2023). Contextual gaps in machine learning for mental illness prediction: The case of diagnostic disclosures. *Proceedings of the ACM on Human-Computer Interaction*, *7*(2), 1-27. https://doi.org/10.1145/3610181

Chang, C. C. H., Yang, H. C., Tang, G., & Ganguli, M. (2009). Minimizing attrition bias: A longitudinal study of depressive symptoms in an elderly cohort. *International Psychogeriatrics*, *21*(5), 869-878. https://doi.org/10.1017/S104161020900876X

Chen, E. & Wojcik, S. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. https://doi.org/10.1037/met0000111

Cho, G., Yim, J., Choi, Y., Ko, J. & Lee, S.-H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investigation*, *16*(4), 262–269. https://doi.org/10.30773/pi.2018.12.21.2

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Chung, J. & Teo, J. (2022). Mental health prediction using machine learning: Taxonomy, applications, and challenges. *Applied Computational Intelligence and Soft Computing*, *2022*(1), 9970363. https://doi.org/10.1155/2022/9970363

Cieslak, D.A., Chawla, N.V. (2008). Learning Decision Trees for Unbalanced Data. In: Daelemans, W., Goethals, B., Morik, K. (Eds.), *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. Lecture Notes in Computer Science* (pp. 241-256). Heidelberg: Springer. https://doi.org/10.1007/978-3-540-87479-9_34

Colodro-Conde, L., Couvy-Duchesne, B., Zhu, G., Coventry, W. L., Byrne, E. M., Gordon, S. et al. (2018). A direct test of the diathesis–stress model for depression. *Molecular Psychiatry*, *23*(7), 1590–1596. https://doi.org/10.1038/mp.2017.130

Conover, W. J. (1999). *Practical nonparametric statistics*. New Jersey: John Wiley & Sons.

Cross-Denny, B. & Robinson, M. A. (2017). Using the social determinants of health as a framework to examine and address predictors of depression in later life. *Ageing International*, *42*(4), 393–412. https://doi.org/10.1007/s12126-017-9278-6

Dao, B., Nguyen, T., Phung, D. & Venkatesh, S. (2014). Effect of mood, social connectivity and age in online depression community via topic and linguistic analysis. In B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali & Y. Zhang (Eds.), *Web Information Systems Engineering – WISE 2014* (pp. 398–407). Cham: Springer. https://doi.org/10.1007/978-3-319-11749-2_30

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine learning* (pp. 233-240). New York: ACM. https://doi.org/10.1145/1143844.1143874

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E. et al. (2022). Digital health tools for the passive monitoring of depression: A systematic review of methods. *npj Digital Medicine*, *5*(1), 1–14. https://doi.org/10.1038/s41746-021- 00548-8

Deci, E. L. & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, *19*(2), 109–134. https://doi.org/10.1016/0092-6566(85)90023-6

Décieux, J., Mergener, A., Neufang, K. & Sischka, P. (2015). Implementation of the forced answering option within online surveys: Do higher item response rates come at the expense of participation and answer quality? *Psihologija*, *48*(4), 311–326. https://doi.org/10.2298/PSI1504311D

Degenhardt, F., Seifert, S. & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, *20*(2), 492–503. https://doi.org/10.1093/bib/bbx124

Demerouti, E., Bakker, A. B., Nachreiner, F. & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology*, *86*(3), 499–512. https://doi.org/10.1037/0021-9010.86.3.499

Denton, B. T. (2018). Optimization of sequential decision making for chronic diseases: From data to decisions (INFORMS TutORials in operations research). *Recent Advances in Optimization and Modeling of Contemporary Problems,* 316–348. https://doi.org/10.1287/educ.2018.0184

Deutsche Sporthochschule Köln (2025). *Das ActIv-Projekt.* Zugriff am 02.06.2025. Abgerufen unter: https://www.dshs-koeln.de/activ-projekt/forschung/activ-projekt/

Devine, J., Fliege, H., Kocalevent, R., Mierke, A., Klapp, B. F. & Rose, M. (2016). Evaluation of computerized adaptive tests (CATs) for longitudinal monitoring of depression, anxiety, and stress reactions. *Journal of Affective Disorders*, *190*, 846–853. https://doi.org/10.1016/j.jad.2014.10.063

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27*(3), 326–327. https://doi.org/10.1145/212094.212114

FORS. (2023). *Documentation | Swiss Household-Panel (SHP).* Zugriff am 24.02.2025. Abgerufen unter: https://forscenter.ch/projects/swiss-household-panel/documentation/

Domingos, P. (1999). The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, *3*, 409-425. https://doi.org/10.1023/A:1009868929893

Eder, J., Dom, G., Gorwood, P., Kärkkäinen, H., Decraene, A., Kumpf, U. et al. (2023). Improving mental health care in depression: A call for action. *European Psychiatry*, *66*(1), e65. https://doi.org/10.1192/j.eurpsy.2023.2434

Ehlert, U. (2014). Psychoendokrinologie: Trends und ihr Bezug zur Praxis. *Verhaltenstherapie*, *24*(2), 125–128. https://doi.org/10.1159/000362915

Eid, R. S., Gobinath, A. R. & Galea, L. A. M. (2019). Sex differences in depression: Insights from clinical and preclinical studies. *Progress in Neurobiology*, *176*, 86–102. https://doi.org/10.1016/j.pneurobio.2019.01.006

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (pp. 973–978). San Francisco: Morgan Kaufmann.

Elragal, A. & Klischewski, R. (2017). Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *Journal of Big Data*, *4*(1), 1-20. https://doi.org/10.1186/s40537-017-0079-2

Evenson, K. R., Goto, M. M. & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *International Journal of Behavioral Nutrition and Physical Activity*, *12*(1), 1-22. https://doi.org/10.1186/s12966-015-0314-1

Fatima, S., Hussain, A., Amir, S. B., Ahmed, S. H., & Aslam, S. M. H. (2023). XGBoost and random forest algorithms: An in-depth analysis. *Pakistan Journal of Scientific Research*, *3*(1), 26-31. https://doi.org/10.57041/vol3iss1pp26-31

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fiedler, K., Nickel, S., Muehlfriedel, T., & Unkelbach, C. (2001). Is mood congruency an effect of genuine memory or response bias?. *Journal of Experimental Social Psychology*, *37*(3), 201-214. https://doi.org/10.1006/jesp.2000.1442

Flett, G. L., Hewitt, P. L., Blankstein, K. R. & Mosher, S. W. (1995). Perfectionism, life events, and depressive symptoms: A test of a diathesis-stress model. *Current Psychology*, *14*(2), 112–137. https://doi.org/10.1007/BF02686885

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J. & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment*, *189*(7), 1-20. https://doi.org/10.1007/s10661-017-6025-0

Fried, E. I., Bockting, C., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, A. O., ... & Stroebe, M. (2015). From loss to loneliness: The relationship between bereavement and depressive symptoms. *Journal of Abnormal Psychology*, *124*(2), 256-265. https://doi.org/10.1037/abn0000028

Fuller, D., Colwell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B. et al. (2020). Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: Systematic review. *JMIR mHealth and uHealth*, *8*(9), e18694. https://doi.org/10.2196/18694

Gesundheitsförderung Schweiz. (2024). *Job-Stress-Index.* Zugriff am 12.03.2025. Abgerufen unter: https://gesundheitsfoerderung.ch/betriebliches-gesundheitsmanagement/themen-und-publikationen/themen/job-stress-index

Geurts, P., Ernst, D. & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gheshlagh, R. G., Sayehmiri, K., Ebadi, A., Dalvandi, A., Dalvand, S., Maddah, S. B., & Tabrizi, K. N. (2017). The relationship between mental health and resilience: A systematic review and meta-analysis. *Iranian Red Crescent Medical Journal*, *19*(6), 1-9. https://doi.org/10.5812/ircmj.13537

Ghojogh, B. & Crowley, M. (2019). Linear and quadratic discriminant analysis: Tutorial. *arXiv Preprint.* https://doi.org/10.48550/arXiv.1906.02590

Glass, D. C., Lake, C. R., Contrada, R. J., Kehoe, K. & Erlanger, L. R. (1983). Stability of individual differences in physiological responses to stress. *Health Psychology*, *2*(4), 317–341. https://doi.org/10.1037/0278-6133.2.4.317

Gregory, A. M., Caspi, A., Eley, T. C., Moffitt, T. E., O'Connor, T. G. & Poulton, R. (2005). Prospective longitudinal associations between persistent sleep problems in childhood and anxiety and depression disorders in adulthood. *Journal of Abnormal Child Psychology*, *33*(2), 157–163. https://doi.org/10.1007/s10802-005-1824-0

Groot, W., van den Brink, H. M. & van Praag, B. (2007). The compensating income variation of social capital. *Social Indicators Research*, *82*(2), 189–207. https://doi.org/10.1007/s11205-006-9035-9

Hakanen, J. J., Schaufeli, W. B. & Ahola, K. (2008). The job demands-resources model: A three-year cross-lagged study of burnout, depression, commitment, and work engagement. *Work & Stress*, *22*(3), 224–241. https://doi.org/10.1080/02678370802379432

Hakulinen, C., Elovainio, M., Pulkki-Råback, L., Virtanen, M., Kivimäki, M., & Jokela, M. (2015). Personality and depressive symptoms: Individual participant meta-analysis of 10 cohort studies. *Depression and Anxiety*, *32*(7), 461-470. https://doi.org/10.1002/da.22376

Hancock, J. T., Khoshgoftaar, T. M. & Johnson, J. M. (2023). Evaluating classifier performance with highly imbalanced big data. *Journal of Big Data*, *10*(1), 10-42. https://doi.org/10.1186/s40537-023-00724-5

Hankin, B. L., & Abela, J. R. (Eds.). (2005). *Development of psychopathology: A vulnerability-stress perspective*. Kalifornien: Sage Publications.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2017). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85. https://doi.org/10.1007/BF0298580

Hatton, C. M., Paton, L. W., McMillan, D., Cussens, J., Gilbody, S., & Tiffin, P. A. (2019). Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare. *Journal of Affective Disorders*, *246*, 857-860. https://doi.org/10.1016/j.jad.2018.12.095

He, H. & Garcia, E. A. (2009). Learning from imbalanced data. *Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hey, T., Tansley, S. & Tolle, K. (Eds.). (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond: Microsoft Research. Abgerufen unter: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf

Hoffmann, P. (1997). The endorphin hypothesis. In W. P. Morgan (Ed.), *Physical activity and mental health* (pp. 163–178). London: Taylor & Francis. https://doi.org/10.4324/9780203782361

Holt-Lunstad, J., Smith, T. B., Baker, M., Harris, T. & Stephenson, D. (2015). Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspectives on Psychological Science*, *10*(2), 227–237. https://doi.org/10.1177/1745691614568352

Höltge, J., Theron, L., & Ungar, M. (2022). A multisystemic perspective on the temporal interplay between adolescent depression and resilience-supporting individual and social resources. *Journal of Affective Disorders*, *297*(3), 225-232. https://doi.org/10.1016/j.jad.2021.10.030

Hopcroft, R. L. & Bradley, D. B. (2007). The sex difference in depression across 29 countries. *Social Forces*, *85*(4), 1483–1507. https://doi.org/10.1353/sof.2007.0071

Hoyer, J. & Knappe, S. (Hrsg.). (2020). *Klinische Psychologie & Psychotherapie*. Heidelberg: Springer. https://doi.org/10.1007/978-3-662-61814-1

Huberty, C. J. (1975). Discriminant analysis. *Review of Educational Research*, *45*(4), 543– 598. https://doi.org/10.3102/00346543045004543

Schmidt, J. C. (2022). Von den Verfahren und Methoden zu den Objekten: Schwache Regelhaftigkeit nutzen – komplexe Objektsysteme werden zugänglich. In C. F. Gethmann, P. Buxmann, J. Distelrath, B. G. Humm, S. Lingner, V. Nitsch & J. C. Schmidt (Hrsg.), *Künstliche Intelligenz in der Forschung: Neue Möglichkeiten und Herausforderungen für die Wissenschaft* (S. 99–104). Heidelberg: Springer. https://doi.org/10.1007/978-3-662-63449-3_2

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning* (Vol. 8). New York: Springer.

Jacoby, R., Barsky, K. G., Porat, T., Harel, S., Miller, T. H. & Goldzweig, G. (2021). Individual stress response patterns: Preliminary findings and possible implications. *PLOS ONE*, *16*(8), e0255889. https://doi.org/10.1371/journal.pone.0255889

Jäger, S., Allhorn, A. & Biessmann, F. (2021). A benchmark for data imputation methods. *Frontiers in Big Data*, *4, 693674*. https://doi.org/10.3389/fdata.2021.693674

Johnston, R. (2023). Discriminant analysis. In F. Maggino (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 1845–1847). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-17299-1_750

Kaplan, G. A., Roberts, R. E., Camacho, T. C. & Coyne, J. C. (1987). Psychosocial predictors of depression: Prospective evidence from the human population laboratory studies. *American Journal of Epidemiology*, *125*(2), 206–220. https://doi.org/10.1093/oxfordjournals.aje.a114521

Karasek, R. A. (1979). Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative Science Quarterly*, *24*(2), 285–308. https://doi.org/10.2307/2392498

Keenan, K., Feng, X., Hipwell, A., & Klostermann, S. (2009). Depression begets depression: Comparing the predictive utility of depression and anxiety symptoms to later depression. *Journal of Child Psychology and Psychiatry*, *50*(9), 1167-1175. https://doi.org/10.1111/j.1469-7610.2009.02080.x

Kemp, A. H., Arias, J. A. & Fisher, Z. (2017). Social ties, health and wellbeing: A literature review and model. In A. Ibáñez, L. Sedeño & A.M. García (Eds.), *Neuroscience and social science: The missing link* (pp. 397–427). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68421-5_17

Kim, S.-Y., Park, J.-H., Lee, M. Y., Oh, K.-S., Shin, D.-W. & Shin, Y.-C. (2019). Physical activity and the prevention of depression: A cohort study. *General Hospital Psychiatry*, *60*, 90–97. https://doi.org/10.1016/j.genhosppsych.2019.07.010

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Klein, E. M., Brähler, E., Dreier, M., Reinecke, L., Müller, K. W., Schmutzer, G., ... & Beutel, M. E. (2016). The german version of the perceived stress scale–psychometric characteristics in a representative german community sample. *BMC psychiatry*, *16*(1), 1-10. https://doi.org/10.1186/s12888-016-0875-9

Kokanovic, R., Bendelow, G. & Philip, B. (2013). Depression: The ambivalence of diagnosis. *Sociology of Health & Illness*, *35*(3), 377–390. https://doi.org/10.1111/j.1467- 9566.2012.01486.x

Kraus, C., Kadriu, B., Lanzenberger, R., Zarate Jr., C. A. & Kasper, S. (2019). Prognosis and improved outcomes in major depression: A review. *Translational Psychiatry*, *9*(1), 1–17. https://doi.org/10.1038/s41398-019-0460-3

Kroenke, K., Spitzer, R. L. & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525- 1497.2001.016009606.x

Kuckartz, U., Rädiker, S., Ebert, T., & Schehl, J. (2013). *Statistik: eine verständliche Einführung*. Heidelberg: Springer.

Kursa, M. B. & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Lehmann, M., Foster, C., & Keul, J. (1993). Overtraining in endurance athletes: A brief review. *Medicine & Science in Sports & Exercise*. *25*(7), 854-862. https://doi.org/10.1249/00005768-199307000

Lépine, J.-P. & Briley, M. (2011). The increasing burden of depression. *Neuropsychiatric Disease and Treatment*, *7*(1), 3–7. https://doi.org/10.2147/NDT.S19617

Lerner, D. & Henke, R. M. (2008). What does research tell us about depression, job performance, and work productivity? *Journal of Occupational and Environmental Medicine*, *50*(4), 401 - 410. https://doi.org/10.1097/JOM.0b013e31816bae50

Letourneau, N. L., Duffett-Leger, L., Levac, L., Watson, B., & Young-Morris, C. (2013). Socioeconomic status and child development: A meta-analysis. *Journal of Emotional and Behavioral Disorders*, *21*(3), 211-224. https://doi.org/10.1177/1063426611421007

Levis, B., Benedetti, A., & Thombs, B. D. (2019). Accuracy of patient health questionnaire- 9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *BMJ, 365*, l1476. https://doi.org/10.1136/bmj.l1476

Liang, Y., Zheng, X. & Zeng, D. D. (2019). A survey on big data-driven digital phenotyping of mental health. *Information Fusion*, *52*, 290–307. https://doi.org/10.1016/j.inffus.2019.04.001

Linardon, J., Messer, M., Goldberg, S. B., & Fuller-Tyszkiewicz, M. (2024). The efficacy of mindfulness apps on symptoms of depression and anxiety: An updated meta-analysis of randomized controlled trials. *Clinical Psychology Review*, *107*, 102370. https://doi.org/10.1016/j.cpr.2023.102370

Liu, S., Roemer, F., Ge, Y., Bedrick, E. J., Li, Z.-M., Guermazi, A. et al. (2023). Comparison of evaluation metrics of deep learning for imbalanced imaging data in osteoarthritis studies. *Osteoarthritis and Cartilage*, *31*(9), 1242–1248. https://doi.org/10.1016/j.joca.2023.05.006

Liu, Y., Zhang, N., Bao, G., Huang, Y., Ji, B., Wu, Y. et al. (2019). Predictors of depressive symptoms in college students: A systematic review and meta-analysis of cohort studies. *Journal of Affective Disorders*, *244*, 196–208. https://doi.org/10.1016/j.jad.2018.10.084

Loe, B. S., Stillwell, D. & Gibbons, C. (2017). Computerized adaptive testing provides reliable and efficient depression measurement using the CES-D Scale. *Journal of Medical Internet Research*, *19*(9), e7453. https://doi.org/10.2196/jmir.7453

Lovato, N. & Gradisar, M. (2014). A meta-analysis and model of the relationship between sleep and depression in adolescents: Recommendations for future research and clinical practice. *Sleep Medicine Reviews*, *18*(6), 521–529. https://doi.org/10.1016/j.smrv.2014.03.006

Lukmanji, A., Williams, J. V. A., Bulloch, A. G. M., Bhattarai, A. & Patten, S. B. (2019). Seasonal variation in symptoms of depression: A Canadian population based study. *Journal of Affective Disorders*, *255*, 142–149. https://doi.org/10.1016/j.jad.2019.05.040

Lukmanji, A., Williams, J. V. A., Bulloch, A. G. M. & Patten, S. B. (2020). Seasonal variation in specific depressive symptoms: A population based study. *Journal of Affective Disorders*, *261*, 153–159. https://doi.org/10.1016/j.jad.2019.10.003

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1-18. https://doi.org/10.1016/S0169-7439(99)00047-7

Makhmutova, M., Kainkaryam, R., Ferreira, M., Min, J., Jaggi, M. & Clay, I. (2021a). PSYCHE-D: Predicting change in depression severity using person-generated health data (DATASET). Zenodo. https://doi.org/10.5281/zenodo.5085146

Makhmutova, M., Kainkaryam, R., Ferreira, M., Min, J., Jaggi, M., & Clay, I. (2021b). Prediction of self-reported depression scores using person-generated health data from a virtual 1-year mental health observational study. In *Proceedings of the 2021 Workshop on Future of Digital Biomarkers* (pp. 4-11). https://doi.org/10.1145/3469266.3469878

Mammen, G. & Faulkner, G. (2013). Physical activity and the prevention of depression: A systematic review of prospective studies. *American Journal of Preventive Medicine*, *45*(5), 649–657. https://doi.org/10.1016/j.amepre.2013.08.001

Manea, L., Gilbody, S. & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the patient health questionnaire (PHQ-9): A meta-analysis. *CMAJ*, *184*(3), 191–196. https://doi.org/10.1503/cmaj.110829

Martinez-Martin, N., Greely, H. T. & Cho, M. K. (2021). Ethical development of digital phenotyping tools for mental health applications: Delphi study. *JMIR mHealth and uHealth*, *9*(7), e27343. https://doi.org/10.2196/27343

Matt, G. E., Vázquez, C., & Campbell, W. K. (1992). Mood-congruent recall of affectively toned stimuli: A meta-analytic review. *Clinical Psychology Review*, *12*(2), 227-255. https://doi.org/10.1016/0272-7358(92)90116-P

McEwen, B. S. (1998a). Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences*, *840*(1), 33-44. https://doi.org/10.1111/j.1749-6632.1998.tb09546.x

McEwen, B. S. (1998b). Protective and damaging effects of stress mediators. *New England Journal of Medicine*, *338*(3), 171-179. https://doi.org/10.1056/nejm199801153380307

Moore, D. S., McCabe, G. P. & Craig, B. A. (2014). *Introduction to the practice of statistics*. San Francisco: W. H. Freeman & Co. Ltd.

Moriarty, A. S., Gilbody, S., McMillan, D. & Manea, L. (2015). Screening and case finding for major depressive disorder using the patient health questionnaire (PHQ-9): A meta-analysis. *General Hospital Psychiatry*, *37*(6), 567–576. https://doi.org/10.1016/j.genhosppsych.2015.06.012

Morris, B. H., Bylsma, L. M. & Rottenberg, J. (2009). Does emotion predict the course of major depressive disorder? A review of prospective studies. *British Journal of Clinical Psychology*, *48*(3), 255–273. https://doi.org/10.1348/014466508X396549

Musliner, K. L., Munk-Olsen, T., Eaton, W. W. & Zandi, P. P. (2016). Heterogeneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes. *Journal of Affective Disorders*, *192*, 199–211. https://doi.org/10.1016/j.jad.2015.12.030

National Institute of Mental Health (NIMH). (2024). *Depression*. Zugriff am 18.09.2024. Abgerufen unter: https://www.nimh.nih.gov/health/topics/depression

Nazareth, I. & Kendrick, T. (2014). Prevention and early intervention in depression and anxiety disorders. In P. Byrne & A. Rosen (Eds.), *Early Intervention in Psychiatry: EI of Nearly Everything for Better Mental Health* (pp. 185-200). New Jersey: Wiley. https://eprints.soton.ac.uk/373520/

Nease, D. E., & Malouin, J. M. (2003). Depression screening: a practical strategy. *Journal of Family Practice*, *52*(2), 118-126.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*, 625-632. https://doi.org/10.1007/s10459-010-9222-y

Ong, C. W., Pierce, B. G., Klein, K. P., Hudson, C. C., Beard, C. & Björgvinsson, T. (2022). Longitudinal measurement invariance of the PHQ-9 and GAD-7. *Assessment*, *29*(8), 1901–1916. https://doi.org/10.1177/10731911211035833

Onnela, J.-P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology*, *46*(1), 45–54. https://doi.org/10.1038/s41386-020-0771-3

Ozenne, B., Subtil, F. & Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, *68*(8), 855–859. https://doi.org/10.1016/j.jclinepi.2015.02.010

Peake, J. M., Kerr, G. & Sullivan, J. P. (2018). A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in Physiology*, *9*. 743. https://doi.org/10.3389/fphys.2018.00743

Pearce, M., Garcia, L., Abbas, A., Strain, T., Schuch, F. B., Golubic, R. et al. (2022). Association between physical activity and risk of depression: A systematic review and meta-analysis. *JAMA Psychiatry*, *79*(6), 550–559. https://doi.org/10.1001/jamapsychiatry.2022.0609

Perna, G., Riva, A., Defillo, A., Sangiorgio, E., Nobile, M. & Caldirola, D. (2020). Heart rate variability: Can it serve as a marker of mental health resilience?. *Journal of Affective Disorders*, *263*, 754–761. https://doi.org/10.1016/j.jad.2019.10.017

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

Pfefferbaum, B. & North, C. S. (2020). Mental health and the covid-19 pandemic. *New England Journal of Medicine*, *383*(6), 510–512. https://doi.org/10.1056/NEJMp2008017

Pinel, J., Barnes, S., Pauli, P. (2018). *Biopsychologie*. London: Pearson.

Poole, D. N., Liao, S., Larson, E., Hedt-Gauthier, B., Raymond, N. A., Bärnighausen, T. et al. (2020). Sequential screening for depression in humanitarian emergencies: A validation study of the patient health questionnaire among syrian refugees. *Annals of General Psychiatry*, *19*(1), 1 - 10. https://doi.org/10.1186/s12991-020-0259-x

Presson, P. K., & Benassi, V. A. (1996). Locus of control orientation and depressive symptomatology: A meta-analysis. *Journal of Social Behavior and Personality*, *11*(1), 201-212.

Probst, P., Wright, M. N. & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Provost, F. (2000). Machine learning from imbalanced data sets 101. *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets* (pp. 68–72). California: AAAI Press. https://pages.stern.nyu.edu/fprovost/Papers/skew.PDF

Ramamoorthy, S., & Cidlowski, J. A. (2016). Corticosteroids-mechanisms of action in health and disease. *Rheumatic Diseases Clinics of North America*, *42*(1), 15-31. https://doi.org/10.1016/j.rdc.2015.08.002

Ramayah, T., Ahmad, N. H., Halim, H. A., Zainal, S. R. M. & Lo, M.-C. (2010). Discriminant analysis: An illustrated example. *African Journal of Business Management, 4*(9), 1654 - 1667. https://doi.org/10.5897/AJBM.9000211

Ramesh, A., Nayak, T., Beestrum, M., Quer, G. & Pandit, J. A. (2023). Heart rate variability in psychiatric disorders: A systematic review. *Neuropsychiatric Disease and Treatment*, *19*(20), 2217–2239. https://doi.org/10.2147/NDT.S429592

Rao, U., Hammen, C. L. & Poland, R. E. (2009). Risk markers for depression in adolescents: Sleep and HPA measures. *Neuropsychopharmacology*, *34*(8), 1936–1945. https://doi.org/10.1038/npp.2009.27

Razzaghi, T., Roderick, O., Safro, I., & Marko, N. (2016). Multilevel weighted support vector machine for classification on healthcare data with missing values. *PloS ONE*, *11*(5), e0155119. https://doi.org/10.1371/journal.pone.0155119

Reinders, H. (2011). Fragebogen. In: Reinders, H., Ditton, H., Gräsel, C., Gniewosz, B. (Hrsg.) Empirische Bildungsforschung (S. 53 – 67). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93015-2_4

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, *3*, 1371–1382. http://doi.org/10.1162/153244303322753715

Richardson, K. M. & Rothstein, H. R. (2008). Effects of occupational stress management intervention programs: A meta-analysis. *Journal of Occupational Health Psychology*, *13*(1), 69–93. Washington: Educational Publishing Foundation. https://doi.org/10.1037/1076-8998.13.1.69

Rimal, Y., Sharma, N. & Alsadoon, A. (2024). The accuracy of machine learning models relies on hyperparameter tuning: Student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. *Multimedia Tools and Applications*, *83*(30), 74349–74364. https://doi.org/10.1007/s11042-024-18426-2

Roddy, M. K., Pfammatter, A. F., & Mayberry, L. S. (2023). Optimizing adaptive stepped-care interventions to change adults' health behaviors: A systematic review. *Journal of Clinical and Translational Science*, *7*(1), e190. https://doi.org/10.1017/cts.2023.618

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Roijers, D. M., Vamplew, P., Whiteson, S. & Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, *48*, 67–113. https://doi.org/10.1613/jair.3987

Rotari, M., & Kulahci, M. (2024). Variable selection wrapper in presence of correlated input variables for random forest models. *Quality and Reliability Engineering International*, *40*(1), 297-312. https://doi.org/10.1002/qre.3398

Rude, S. S., Wenzlaff, R. M., Gibbs, B., Vane, J., & Whitney, T. (2002). Negative processing biases predict subsequent depressive symptoms. *Cognition & Emotion*, *16*(3), 423-440. https://doi.org/10.1080/02699930143000554

Ruız, L. C., & Kovacs, T. (2021). Biometric screenings: The route to occupational safety and health. *Journal of Scientific Perspectives*, *5*(1), 25-35. https://doi.org/10.26900/jsp.5.1.03

Ryan, R. M., Williams, G. C., Patrick, H. & Deci, E. L. (2009). Self-determination theory and physical activity: The dynamics of motivation in development and wellness. *Hellenic Journal of Psychology*, *6*(2), 107–124. Thessaloniki: Psychological Society of Northern Greece.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Santa Maria, A., Wörfel, F., Wolter, C., Gusy, B., Rotter, M., Stark, S. et al. (2018). The role of job demands and job resources in the development of emotional exhaustion, depression, and anxiety among police officers. *Police Quarterly*, *21*(1), 109–134. https://doi.org/10.1177/1098611117743957

Santini, Z. I., Koyanagi, A., Tyrovolas, S., Mason, C. & Haro, J. M. (2015). The association between social relationships and depression: A systematic review. *Journal of Affective Disorders*, *175*(1), 53–65. https://doi.org/10.1016/j.jad.2014.12.049

Schaufeli, W. B. & Taris, T. W. (2014). A critical review of the job demands-resources model: Implications for improving work and health. In G.F. Bauer & O. Hämmig (Eds.), *Bridging occupational, organizational and public health: A transdisciplinary approach* (pp. 43–68). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-5640-3_4

Schupp, J. & Wolf, C. (Hrsg.). (2015). *Nonresponse Bias: Qualitätssicherung sozialwissenschaftlicher Umfragen*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-10459-7

SHP Group. (2024). Living in Switzerland Waves 1-25 + Long file + Covid 19 data (Version 9.0) [Data set]. FORS. https://doi.org/10.48573/swnc-bn46

Schwyck, M., Lieberman, M. & Parkinson, C. (2020). Multivoxel pattern analysis in fMRI: A practical introduction for social and affective neuroscientists. *Social Cognitive and Affective Neuroscience*, *15*(4), 487 - 509. https://doi.org/10.1093/scan/nsaa057

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, *60*, 144–162. https://doi.org/10.1051/proc/201760144

Sekhar, Ch. R., Minal & Madhu, E. (2016). Mode choice analysis using random forest decision trees. *Transportation Research Procedia*, *17*, 644–652. https://doi.org/10.1016/j.trpro.2016.11.119

Shin, D., Lee, K. J., Adeluwa, T. & Hur, J. (2020). Machine learning-based predictive modeling of postpartum depression. *Journal of Clinical Medicine*, *9*(9), 2899. https://doi.org/10.3390/jcm9092899

Simjanoski, M., Ballester, P. L., da Mota, J. C., De Boni, R. B., Balanzá-Martínez, V., Atienza-Carbonell, B. et al. (2022). Lifestyle predictors of depression and anxiety during covid-19: A machine learning approach. *Trends in Psychiatry and Psychotherapy*, *44*, e20210365. https://doi.org/10.47626/2237-6089-2021-0365

Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Stewart, S. M., Kennard, B. D., Lee, P. W. H., Hughes, C. W., Mayes, T. L., Emslie, G. J. et al. (2004). A cross-cultural investigation of cognitions and depressive symptoms in adolescents. *Journal of Abnormal Psychology*, *113*(2), 248–257. https://doi.org/10.1037/0021-843X.113.2.248

Strakowski, S. M. (2012). The complexities of depression. *Current Psychiatry Reports*, *14*(6), 608–609. https://doi.org/10.1007/s11920-012-0330-7

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(307), 1-11. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 1-21. https://doi.org/10.1186/1471-2105-8-25

Su, C., Ju, S., Liu, Y. & Yu, Z. (2015). Improving random forest and rotation forest for highly imbalanced datasets. *Intelligent Data Analysis*, *19*(6), 1409–1432. https://doi.org/10.3233/IDA-150789

Tan, Q., Cai, Y., Li, Q., Zhang, Y. & Tu, D. (2018). Development and validation of an item bank for depression screening in the chinese population using computer adaptive testing: A simulation study. *Frontiers in Psychology*, *9,* 1225. https://doi.org/10.3389/fpsyg.2018.01225

Tesarova, S., Pekacek, O. & Porrovecchio, A. (2023). Predictors of depression: Lifestyle choices during the pandemic. *Frontiers in Psychology*, *14,* 1194270. https://doi.org/10.3389/fpsyg.2023.1194270

Tillmann, R., Voorpostel, M., Antal, E., Dasoki, N., Klaas, H., Kuhn, U. et al. (2022). The swiss household panel (SHP). *Jahrbücher für Nationalökonomie und Statistik*, *242*(3), 403– 420. https://doi.org/10.1515/jbnst-2021-0039

Torterolo, P., Scorza, C., Lagos, P., Urbanavicius, J., Benedetto, L., Pascovich, C. et al. (2015). Melanin-concentrating hormone (MCH): Role in REM sleep and depression. *Frontiers in Neuroscience*, *9,* 475. https://doi.org/10.3389/fnins.2015.00475

Tourangeau, R., Rips, L. J., & Rasinski, K. (Eds.). (2000). *The psychology of survey response.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Trapeznikov, K. & Saligrama, V. (2013). Supervised sequential classification under budget constraints. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (pp. 581–589). https://proceedings.mlr.press/v31/trapeznikov13a.html

Tsaur, S.-H., Liang, Y.-W. & Hsu, H.-J. (2012). A multidimensional measurement of work- leisure conflict. *Leisure Sciences*, *34*(5), 395–416. https://doi.org/10.1080/01490400.2012.714701

Tsaur, S.-H. & Yen, C.-H. (2018). Work–leisure conflict and its consequences: Do generational differences matter? *Tourism Management*, *69*(4), 121–131. https://doi.org/10.1016/j.tour- man.2018.05.011

Tsuno, N., Besset, A. & Ritchie, K. (2005). Sleep and depression. *The Journal of Clinical Psychiatry*, *66*(10), 1254–1269. https://doi.org/10.4088/jcp.v66n1008

Universität Bern. (2022). *POSEIDON-298* [Gitlab Repository]. Zugriff am 18.9.2024. Abgerufen unter: https://git.upd.unibe.ch/openscience/poseidon

US Preventive Services Task Force, (2002). Screening for depression: recommendations and rationale. *Annals of Internal Medicine*, *136*(10), 760-764. https://doi.org/10.7326/0003-4819-136-10-200205210-00012

US Preventive Services Task Force, (2023). *Depression and Suicide Risk in Adults: Screening.* Zugriff am 02.06.2025. Abgerufen unter: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/screening-depression-suicide-risk-adults

Valavi, R., Elith, J., Lahoz-Monfort, J. J. & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, *44*(12), 1731–1742. https://doi.org/10.1111/ecog.05615

van der Ploeg, T., Austin, P. C. & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*(1), 1 - 13. https://doi.org/10.1186/1471-2288-14-137

Van Lissa, C. J. (2023). Developmental data science: How machine learning can advance theory formation in developmental psychology. *Infant and Child Development*, *32*(6), e2370. https://doi.org/10.1002/icd.2370

Vindegaard, N. & Benros, M. E. (2020). Covid-19 pandemic and mental health consequences: Systematic review of the current evidence. *Brain, Behavior, and Immunity*, *11*(2), 531– 542. https://doi.org/10.1016/j.bbi.2020.05.048

Voorpostel, M., Tillmann, R., Lebert, F., Kuhn, U., Lipps, O., Ryser, V.A., Antal E., Dasoki, N., Janssen, C., & Wernli B. (2024). *Swiss Household Panel User Guide (1999-2022), Wave 24, February 2024.* Lausanne: FORS.

Wainer, J. & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, *182*, 115222. https://doi.org/10.1016/j.eswa.2021.115222

Wanderley Espinola, C., Gomes, J. C., Mônica Silva Pereira, J. & dos Santos, W. P. (2022). Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: An exploratory study. *Research on Biomedical Engineering*, *38*(3), 813–829. https://doi.org/10.1007/s42600-022-00222-2

Watson, D., Clark, L. A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Weaverdyck, M. E., Lieberman, M. D. & Parkinson, C. (2020). Tools of the trade multivoxel pattern analysis in fMRI: A practical introduction for social and affective neuroscientists. *Social Cognitive and Affective Neuroscience*, *15*(4), 487–509. https://doi.org/10.1093/scan/nsaa057

Weissman, M. M., Bland, R., Joyce, P. R., Newman, S., Wells, J. E. & Wittchen, H.-U. (1993). Sex differences in rates of depression: Cross-national perspectives (special issue toward a new psychobiology of depression in women). *Journal of Affective Disorders*, *29*(2), 77–84. https://doi.org/10.1016/0165-0327(93)90025-F

Whalen, D. J., Gilbert, K. E., Barch, D. M., Luby, J. L. & Belden, A. C. (2017). Variation in common preschool sleep problems as an early predictor for depression and anxiety symptom severity across time. *Journal of Child Psychology and Psychiatry*, *58*(2), 151–159. https://doi.org/10.1111/jcpp.12639

Williams, S. Z., Chung, G. S., & Muennig, P. A. (2017). Undiagnosed depression: A community diagnosis. *SSM-Population Health*, *3*, 633-638. https://doi.org/10.1016/j.ssmph.2017.07.012

de Winter, A. F., Oldehinkel, A. J., Veenstra, R., Brunnekreef, J. A., Verhulst, F. C. & Ormel, J. (2005). Evaluation of non-response bias in mental health determinants and outcomes in a large sample of pre-adolescents. *European Journal of Epidemiology*, *20*(2), 173–181. https://doi.org/10.1007/s10654-004-4948-6

Wong, E. F., Accortt, E. E., Choi, S. W. & Bright, T. J. (2025). Computer adaptive testing strategies for the edinburgh postnatal depression scale (EPDS). *Archives of Women's Mental Health*, *28,* 1 - 10. https://doi.org/10.1007/s00737-025-01562-5

World Health Organization (WHO) (2017). *Depression and other common mental disoders.* Zugriff am 17.02.2025. Abgerufen unter: https://www.who.int/publications/i/item/depres- sion-global-health-estimates

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Eds.), *Methodology of longitudinal surveys* (pp. 157–181). New Jersey: Wiley. http://doi.org/10.1002/9780470743874.ch10

World Health Organization (WHO). (2024). *International classification of diseases, 11th revision (ICD-11)*. Zugriff am 17.02.2025. Abgerufen unter: https://icd.who.int/en

Wyss, P., Ginsbourger, D., Shou, H., Davatzikos, C., Klöppel, S. & Abdulkadir, A. (2023). Adaptive data-driven selection of sequences of biological and cognitive markers in pre-clinical diagnosis of dementia. *Scientific Reports*, *13*(1), 6406. https://doi.org/10.1038/s41598-023-32867-z

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3

Zetsche, U., & Joormann, J. (2011). Components of interference control predict depressive symptoms and rumination cross-sectionally and at six months follow-up. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(1), 65-73. https://doi.org/10.1016/j.jbtep.2010.06.001

Zhai, L., Zhang, H. & Zhang, D. (2015). Sleep duration and depression among adults: A meta-Analysis of prospective studies. *Depression and Anxiety*, *32*(9), 664–670. https://doi.org/10.1002/da.22386

Zhang, J., & Saju, C. (2023). A systematic review of randomised controlled trials with adaptive and traditional group sequential designs–applications in cardiovascular clinical trials. *BMC Medical Research Methodology*, *23*(1), 200. https://doi.org/10.1186/s12874-023-02024-1

Zhang, W., Liu, H., Silenzio, V. M. B., Qiu, P. & Gong, W. (2020). Machine learning models for the prediction of postpartum depression: Application and comparison based on a cohort study. *JMIR Medical Informatics*, *8*(4), e15516. https://doi.org/10.2196/15516

Zhang, X. C., Woud, M. L., Becker, E. S. & Margraf, J. (2018). Do health-related factors predict major depression? A longitudinal epidemiologic study. *Clinical Psychology & Psychotherapy*, *25*(3), 378–387. https://doi.org/10.1002/cpp.2171

Zhang, X., Jeske, D. R., Li, J. & Wong, V. (2016). A sequential logistic regression classifier based on mixed effects with applications to longitudinal data. *Computational Statistics & Data Analysis*, *94*, 238–249. https://doi.org/10.1016/j.csda.2015.08.009

Zubin, J. & Spring, B. (1977). Vulnerability: A new view of schizophrenia. *Journal of Abnormal Psychology*, *86*(2), 103–126. https://doi.org/10.1037/0021-843X.86.2.103

# Verzeichnis der verwendeten Hilfsmittel

| Assistenzsystem | Teile / Stelle(n) in der Arbeit | Einsatz |
|---|---|---|
| *ChatGPT* | *Theorieteil* | *Suchen von Arbeiten in denen bekannte Theorien zum ersten Mal genannt wurden/ihren Ursprung haben.* |
| *ChatGPT* | *Resultate* | *Überprüfen des selbstständig geschriebenen Codes in R auf Fehler und Nachfragen bei einfachen Codes und zu Informationen zu Paketen. Hilfe/Nachfrage bei Fehlermeldungen in R.* |
| *DeepL* | *Theorie* | *Übersetzen von Teilen von komplizierten Arbeiten vom Englischen ins Deutsche (insbesondere bei Theorie zu statistischen Verfahren).* |
| *Zotero* | *Ganzer Bericht* | *Managen von Literatur und Erstellen der ersten Version des Literaturverzeichnisses (danach eigenständig überarbeitet).* |
| *Word* | *Ganzer Bericht* | *Rechtschreib- und Grammatikprüfung.* |
| *LamnguageTool* | *Ganzer Bericht* | *Rechtschreib- und Grammatikprüfung.* |
| *Grammarly* | *Ganzer Bericht* | *Rechtschreib- und Grammatikprüfung.* |
| *Scribbr* | *Ganzer Bericht* | *Plagiatscheck zur Überprüfung unbeabsichtigter Textübereinstimmungen.* |

# Tabellenverzeichnis

Tabelle 1: Operationalisierung der abhängigen Variablen für zukünftige erhöhte depressive Symptomatik ..... 32

Tabelle 2: Verwendete Hyperparameter für das Hyperparametertuning ..... 46

Tabelle 3: Schweizer Haushalt-Panel durch Boruta ausgewählte Variablen ..... 53

Tabelle 4: Pro Fold ausgewählte Hyperparameter und PRAUC Schweizer Haushalt-Panel ... 55

Tabelle 5: PSYCHE-D durch Boruta ausgewählte Variablen..... 56

Tabelle 6: Pro Fold ausgewählte Hyperparameter und PRAUC PSYCHE-D..... 58

Tabelle 7: In der Diskriminanzanalyse erstellten Modelle über Folds hinweg Schweizer Haushalt-Panel..... 60

Tabelle 8: In der Diskriminanzanalyse erstellten Modelle über Folds hinweg PSYCHE-D.... 62

# Abbildungsverzeichnis

Abbildung 1. Vorgehen bis zur Anwendung der sequentiellen Algorithmen. ..... 44
Abbildung 2. BA einmalige, zweifache und sequentielle Messung Schweizer Haushalt-Panel. ..... 63
Abbildung 3. BA einmalige, zweifache und sequentielle Messung PSYCHE-D.. ..... 67
Abbildung 4. Accuracy einmalige, zweifache und sequentielle Messung PSYCHE-D Gewichtung nach inversem Anteil der späteren erhöhten depressiven Symptomatik..... 68
Abbildung 5. Totale Kosten einmalige, zweifache und sequentielle Messung Schweizer-Haushalt-Panel..... 71
Abbildung 6. Totale Kosten einmalige, zweifache und sequentielle Messung PSYCHE-D.... 73
Abbildung 7. Zusammenhang zwischen Geburtsjahr und PHQ-9 Wert PSYCHE-D.. ..... 93